Using Measures of Semantic Relatedness for Word Sense Disambiguation

Siddharth Patwardhan[1], Satanjeev Banerjee[2], and Ted Pedersen[1]

[1] University of Minnesota, Duluth, MN 55812 USA
[2] Carnegie Mellon University, Pittsburgh, PA 15213 USA
patw0006@umn.edu, banerjee+@cs.cmu.edu, tpederse@umn.edu
http://www.d.umn.edu/{~patw0006,~tpederse},
http://www.cs.cmu.edu/~banerjee

Abstract. This paper generalizes the Adapted Lesk Algorithm of Banerjee and Pedersen (2002) to a method of word sense disambiguation based on semantic relatedness. This is possible since Lesk's original algorithm (1986) is based on gloss overlaps which can be viewed as a measure of semantic relatedness. We evaluate a variety of measures of semantic relatedness when applied to word sense disambiguation by carrying out experiments using the English lexical sample data of SENSEVAL-2. We find that the gloss overlaps of Adapted Lesk and the semantic distance measure of Jiang and Conrath (1997) result in the highest accuracy.

1 Introduction

Word sense disambiguation is the process of assigning a meaning to a word based on the context in which it occurs. The most appropriate meaning for a word is selected from a predefined set of possibilities, usually known as a *sense inventory*.

In this paper we present a class of dictionary–based methods that follow from the Adapted Lesk Algorithm of Banerjee and Pedersen [2]. The original Lesk algorithm [9] disambiguates a target word by selecting the sense whose gloss (or definition) has the largest number of words that overlap (or match) with the glosses of neighboring words. Banerjee and Pedersen extend the concept of a gloss overlap to include the glosses of words that are related to the target word and its neighbors according to the concept hierarchies provided in the lexical database WordNet [4]. This paper takes the view that gloss overlaps are just another measure of semantic relatedness, which is a point previously noted by Resnik [13]. In this paper we evaluate several additional measures of semantic relatedness when applied to word sense disambiguation using the general framework provided by the Adapted Lesk Algorithm.

Supervised learning algorithms also assign meanings to words from a sense inventory, but take a very different approach. A human manually annotates examples of a word with tags that indicates the intended sense in each context. These examples become training data for a learning algorithm that induces rules that are then used to assign meanings to other occurrences of the word. In supervised methods, the human uses the information in the dictionary to decide

which sense tag should be assigned to an example, and then a learning algorithm finds clues from the context of that word that allow it to generalize rules of disambiguation. Note that the learning algorithm simply views the sense inventory as a set of categories and that the human has absorbed the information from the dictionary and combined it with their own knowledge of words to manually sense–tag the training examples. The objective of a dictionary–based approach is to provide a disambiguation algorithm with the contents of a dictionary and attempt to make inferences about the meanings of words in context based on that information. Here we extract information about semantic relatedness from the lexical database WordNet (sometimes augmented by corpus statistics) in order to make such inferences.

This paper begins with an overview of the original Lesk algorithm and the adaptation of Banerjee and Pedersen. We review five other measures of semantic relatedness that are included in this study. These include measures by Resnik (1995), Jiang and Conrath (1997), Lin (1997), Leacock and Chodorow (1998), and Hirst and St. Onge (1998). We go on to describe our experimental methodology and results. We close with an analysis and discussion, as well as a brief review of related work.

2 The Lesk Algorithm

The original Lesk algorithm [9] disambiguates a target word by comparing its gloss with those of its surrounding words. The target word is assigned the sense whose gloss has the most overlapping or shared words with the glosses of its neighboring words.

There are two hypotheses that underly this approach. The first is that words that appear together in a sentence can be disambiguated by assigning to them the senses that are most closely related to their neighboring words. This follows from the intuition that words that appear together in a sentence must inevitably be related in some way, since they are normally working together to communicate some idea. The second hypothesis is that related senses can be identified by finding overlapping words in their definitions. The intuition here is equally reasonable, in that words that are related will often be defined using the same words, and in fact may refer to each other in their definitions.

For example, in *The rate of interest at my bank is...* a human reader knows that *bank* refers to a financial institution rather than a river shore, since each of these words has a financial sense. In WordNet the glosses of the financial senses of these three words overlap; the glosses of *interest* and *bank* share *money* and *mortgage*, and the glosses of *interest* and *rate* share *charge*.

The main limitation to this approach is that dictionary glosses are often quite brief, and may not include sufficient vocabulary to identify related senses. Banerjee and Pedersen suggest an adaptation based on the use of WordNet. Rather than simply considering the glosses of the surrounding words in the sentence, the concept hierarchy of WordNet is exploited to allow for glosses of word senses related to the words in the context to be compared as well. In effect,

the glosses of surrounding words in the text are expanded to include glosses of those words to which they are related through relations in WordNet. Pedersen and Banerjee also suggest a variation to the scoring of overlaps such that a match of n consecutive words in two glosses is weighted more heavily than a set of n one word matches.

Suppose that *bark* is the target word and it is surrounded by *dog* and *tail*. The original Lesk algorithm checks for overlaps in the glosses of the senses of *dog* with the glosses of *bark*. Then it checks for overlaps in the glosses of *bark* and *tail*. The sense of *bark* with the maximum number of overlaps with *dog* and *tail* is selected. The adaptation of the Lesk algorithm considers these same overlaps and adds to them the overlaps of the glosses of the senses of concepts that are semantically or lexically related to *dog*, *bark* and *tail* according to WordNet.

3 WordNet

WordNet [4] is a freely–available electronic dictionary of nouns, verbs, adjectives and adverbs that has been developed at Princeton University. It organizes related concepts into *synonym sets* or *synsets*. For example: {*car*, *auto*, *automobile*, *machine*, *motorcar*} is a synset that represents the concept defined by the gloss, *4–wheeled motor vehicle; usually propelled by an internal combustion engine.*

In addition to providing these groups of synonyms to represent a concept, WordNet connects concepts via a variety of relations. This creates a network where related concepts can be (to some extent) identified by their relative distance from each other. The relations provided include *synonymy*, *antonymy*, *is–a*, and *part–of*.

The concept hierarchies in WordNet generally do not cross part of speech boundaries, so semantic and lexical relations are confined to a particular part of speech. For nouns, an *is–a* relation exists between two concepts when one concept *is–a–kind–of* another concept. Such a concept is also known as a *hypernym*. For example, a *car* is a hypernym of *motor vehicle*. An *is–a* hierarchy also exists for verbs, although it represents *is–way–of–doing*, also known as *troponomy*. As an example, *walking* is a troponym of *moving*. Each of these hierarchies has a very general topmost node that is not related to a specific concept. As one traverses down from these topmost nodes the concepts become more specific or topical.

We use WordNet 1.7 which contains nine separate noun hierarchies containing 74,588 concepts joined by 76,226 *is–a* links. In order to allow for paths between all noun concepts in WordNet, we create a root node that subsumes the nine given hierarchies. Verb hierarchies provide less information about relatedness between separate concepts since there are 628 separate hierarchies for the 12,754 verb concepts. While these could all be joined by a root node, the result would be a tree structure that was very wide and would result in many different concepts being located at approximately the same path length from each other.

The measures of semantic relatedness considered in this paper focus on the noun *is–a* hierarchies in WordNet. These are the most developed relations in WordNet, comprising over 70% of the total relations for nouns. Each hierarchy

can be visualized as a tree that has a very general concept associated with a root node and more specific concepts associated with leaves. For example, a root node might represent a concept like *entity* whereas leaf nodes are associated with *carving fork* and *whisk broom*.

Path lengths between concepts have been employed in other networks of concepts to represent semantic relatedness (e.g., [12]). However, this is only appropriate when the path lengths between concepts have a consistent interpretation. This is not the case with WordNet, since concepts higher in a hierarchy are more general than those lower in the hierarchy. Thus, a path of length one between two general concepts can suggest a large difference whereas one between two specific concepts may not. For example, in WordNet *mouse* and *rodent* are separated by a path of length one, which is the same distance that separates *fire iron* and *implement*. The fact that path lengths can be interpreted differently depending on where they occur in WordNet has led to the development of a number of measures based on path lengths that incorporate a variety of correcting factors.

4 Measures of Semantic Relatedness

We make a distinction between *relatedness* and *similarity*, following Budanitsky and Hirst [3]. Semantic similarity is a kind of relatedness between two words that defines a resemblance. Semantic relatedness covers a broader range of relationships between concepts that includes similarity (or difference) as as well other relations such as *is–a–kind–of*, *is–a–part–of*, *is–a–specific–example–of*, *is–the–opposite–of* to name but a few.

There are pairs of words that tend to occur together more often than we'd expect by chance. Sometimes this is indicative of a semantic relationship between the words. Even though these relations are quite diverse, humans can usually judge if a pair of words is more related than another. For example, a human would judge *paper* and *pencil* much more closely related than *car* and *fork*.

It would be useful to assign a value that characterizes the degree to which two words are related. The gloss overlaps discussed previously can be viewed as a very simple mechanism for assigning such values. For example, if two concepts share two words in their respective glosses, they might be considered to be more related than a pair of concepts whose glosses share one word.

What follows is a discussion of a number of measures that have been developed to assign values of semantic relatedness based on their relative position in a concept hierarchy, and possibly augmented by corpus–based information. All of these measures are based on the concept hierarchies as provided by WordNet. Please note that in the rest of this paper *concept* and *word sense* are used somewhat interchangeably, since each concept in WordNet represents a distinct meaning that can be considered a word sense.

4.1 The Leacock–Chodorow Measure

The measure of Leacock and Chodorow [8] is based on the lengths of paths between noun concepts in an *is–a* hierarchy. The shortest path between two

concepts is the one which includes the fewest number of intermediate concepts. This value is scaled by the depth of the hierarchy, where depth is defined as the length of the longest path from a leaf node to the root node of the hierarchy.

Thus, their measure of relatedness is defined as follows:

$$related_{lch}(c_1, c_2) = max[-log(ShortestLength(c_1, c_2)/(2 \cdot D))] \quad (1)$$

$ShortestLength(c_1, c_2)$ is the shortest path length (having minimum number of nodes) between the two concepts and D is the maximum depth of the taxonomy (distance of the farthest node from the root node). Given our scheme of introducing a hypothetical root node that joins all the noun hierarchies, D becomes a constant of 16 for all nouns, meaning that the path length from this root node to the most distant leaf node is 16 in WordNet 1.7.

4.2 The Resnik Measure

Resnik [13] introduces a measure of relatedness based on his formulation of *information content*, which is a value that is assigned to each concept in a hierarchy based on evidence found in a corpus.

Before describing this measure of relatedness we first introduce the notion of information content, which is simply a measure of the specificity of a concept. A concept with a high information content is very specific to a particular topic, while concepts with lower information content are associated with more general, less specific concepts. Thus, *carving fork* has a high information content while *entity* has low information content.

Information content of a concept is estimated by counting the frequency of that concept in a large corpus and thereby determining its probability via a maximum likelihood estimate. According to Resnik, the negative log of this probability determines the information content of the concept:

$$IC(concept) = -log(P(concept)) \quad (2)$$

If sense–tagged text is available, frequency counts of concepts can be attained directly, since each concept will be associated with a unique sense. If sense–tagged text is not available (which is the usual situation) it will be necessary to adopt an alternative counting scheme. Resnik [14] suggests counting the number of occurrences of a word type in a corpus, and then dividing that count by the number of different concepts/senses associated with that word. This value is then assigned to each concept. For example, suppose that the word type *bank* occurs 20 times in a corpus, and that there are two concepts associated with this type in the hierarchy, one for *river bank* and the other for *financial bank*. Each of these concepts would receive a count of 10. If the occurrences of *bank* were sense tagged then the relevant counts could simply be assigned to the appropriate concept.

In our experiments we choose to assign the total count to all the concepts and not divide by the number of possible concepts. Thus we would assign 20 to *river bank* and *financial bank* in the example above. This decision was based on

the observation that by distributing the frequency count over all the concepts associated with a word type we effectively assign a higher relative frequency to those words having fewer senses. This would lead us to estimate a higher probability and therefore assign a lower value of information content to such concepts.

For example, suppose again that *bank* occurs 20 times and that there are two possible underlying concepts. Further suppose that *carving fork* also occurs 20 times but that it only has one associated concept. In the counting scheme of Resnik the two concepts associated with *bank* would have a higher information content than the single concept associated with *carving fork*. However, the term *carving fork* is certainly referring to just one concept, while occurrences of *bank* could be referring to either of the two possible concepts. As such it seems that the information content of *carving fork* should be at least as high as *bank* in this case.

Regardless of how they are counted, the frequency of a concept includes the frequency of all its subordinate concepts since the count we add to a concept is added to its subsuming concept as well. Note that the counts of more specific concepts are added to the more general concepts, but not from the more general to specific. Thus, counts of more specific concepts percolate up to the top of the hierarchy, incrementing the counts of the more general concepts as they proceed upward. As a result, concepts that are higher up in the hierarchy will have higher counts than those at lower more specific levels and have higher probabilities associated with them. Such high probability concepts will have low values of information content since they are associated with more general concepts.

The Resnik measure of semantic similarity [13] uses the information content of concepts along with their positions in the noun *is–a* hierarchies of WordNet to compute a value for the semantic relatedness of the concepts. The principle idea behind his measure of semantic relatedness is that two concepts are semantically related proportional to the amount of information they share in common. The quantity of information common to two concepts is determined by the information content of the lowest concept in the hierarchy that subsumes both the given concepts. This concept is known as the *lowest common subsumer* of the two concepts. Thus, the Resnik measure of similarity is defined as follows:

$$sim_{res}(c_1, c_2) = IC(lcs(c_1, c_2)) \tag{3}$$

We note that this measure does not consider the information content of the concepts themselves, nor does it directly consider the path length. The potential limitation of this approach is that quite a few concepts might have the same least common subsumer, and would have identical values of similarity assigned to them. For example, in WordNet the concept of *vehicle* is the least common subsumer of *jumbo jet*, *tank*, *house trailer*, and *ballistic missile*. Therefore any pair of these concepts would receive the same similarity score.

4.3 The Jiang–Conrath Measure

Jiang and Conrath [7] use information content as defined by Resnik and augment it with a notion of path length between concepts. This results in a hybrid approach to computing semantic relatedness of pairs of concepts. This approach includes the information content of the concepts themselves along with the information content of their lowest common subsumer. The measure is determined by the formula:

$$dist_{jcn}(c_1, c_2) = IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1, c_2)) \tag{4}$$

This formula, however, results in a distance (or measure of unrelatedness) between the two concepts. Concepts that are more related have a lower score than the less related ones. In order to maintain consistency among the measures, we convert this measure of semantic distance into a measure of semantic relatedness via the following:

$$related_{jcn}(c_1, c_2) = \frac{1}{dist_{jcn}(c_1, c_2)} \tag{5}$$

4.4 The Lin Measure

The Lin measure [10] of semantic relatedness of concepts is based on his *Similarity Theorem*. It states that the similarity of two concepts is measured by the ratio of the amount of information needed to state the commonality of the two concepts to the amount of information needed to describe them.

The commonality of two concepts is captured by the information content of their lowest common subsumer and the information content of the two concepts themselves. This measure turns out to be a close cousin of the Jiang–Conrath measure, although they were developed independently:

$$related_{lin}(c_1, c_2) = \frac{2 \times IC(lcs(c_1, c_2))}{IC(c_1) + IC(c_2)} \tag{6}$$

This can be viewed as taking the information content of the intersection of the two concepts (multiplied by 2) and dividing it by their sum, which is analogous to the well–known Dice Coefficient.

4.5 The Hirst–St.Onge Measure

All of the above measures of semantic relatedness consider only the *is–a* relations for nouns in WordNet. Hirst and St. Onge [6] introduce a measure of relatedness that considers many other relations in WordNet and is not restricted to nouns. The measure was originally intended to identify lexical chains, which are a series of words that are related and maintain coherence in a text.

As a result the Hirst–St. Onge measure assigns a relatedness score for word types rather than concepts. In order to make this measure suitable for our purposes, we eliminated one relation (extra strong) from the original formulation that focuses on word types rather than concepts.

This measure classifies all WordNet relations as horizontal, upward, or downward. Upward relations connect more specific concepts to more general ones, while downward relations join more general concepts to more specific ones. For example, *is–a* is an upward relation while *contains* is considered to be a downward relation. Horizontal relations maintain the same level of specificity, where *antonyms* are an example.

The Hirst–St. Onge measure has four levels of relatedness: extra strong, strong, medium strong, and weak. An extra strong relation is based on the surface form of the words and therefore does not apply in our case since we are measuring the relatedness of word senses.

Two words representing the same concept have a strong relation between them. Thus, there is a strong relation between two instances of the same concept. There are two additional scenarios by which a strong relations can exist: First, if the synsets representing the concepts are connected via a horizontal relation then this constitutes a strong relation. Second, if one of the concepts is represented by a compound word and the other concept is represented by a word which is a part of the compound, and if there is any kind of synset relation between the two concepts, then there exists a strong relation between the two concepts.

The medium–strong relation is determined by a set of allowable paths between concepts that are described by Hirst and St. Onge [6]. If such a path exists between two concepts, then we have a medium–strong relation between them. The score or weight for the relation in this case is given by a formula that considers the path length between the concepts and the number of changes in direction of the path:

$$path\_weight = C - path\_length - (k \times number\_of\_changes\_in\_direction) \quad (7)$$

Following Budanitsky and Hirst, we use C = 8 and k = 1. The value of strong relations is defined to be 2*C. Thus, two concepts that exhibit a strong relation will receive a score of 16, two concepts with a medium strong relation will be scored as in the formula above, and two concepts that have no relation will receive a score of zero.

5 Disambiguation using Semantic Relatedness

What follows is a description of the Adapted Lesk Algorithm of Banerjee and Pedersen. It starts by selecting a *window of context* that consists of the target word and some number of content words to the left and right that are known to WordNet. For the experiments in this paper, we use a window of three words, meaning that the glosses of the target word are compared with the glosses of the content words immediately to its left and right. However, if the target word occurs at the beginning or end of the sentence we adjust the window so that the two content words are to the right or left of the target word.

The algorithm identifies candidate senses for each word in the window of context based on the sense inventory in WordNet. If a word in the window is used as part of a compound, then the senses associated with that compound are

the candidates. Otherwise, the candidate senses include those associated with the surface form of the word in the window as well as those of the base form of the word, as determined by the WordNet stemmer.

Most of the measures of semantic relatedness that we employ are intended for use with nouns; the only exceptions are the measures of Hirst–St. Onge and the gloss overlaps of the Adapted Lesk Algorithm. As such we only consider the noun senses of words found in the window of context. We do not part of speech tag the words in the windows of context. Instead, we identify the first word to the left and right of the target word that has a noun form that appears in WordNet, regardless of whether that word is actually used as a noun in that particular context. Our conjecture is that an adjacent verb or adjective that has a noun form will be more related to the target word than will a potentially more distant word that is used as a noun. Thus the window is formed by surrounding words that have noun forms regardless of their actual usage in that context.

After the candidate senses are determined, we measure the relatedness of the candidate senses of the target word to those of the surrounding words in the window of context. From this a score is computed for each sense of the target word that specifies how related it is to the senses of the words in the window of context. While the general framework of Banerjee and Pedersen supports two strategies for computing these scores, we employ their *local* paradigm.

This scoring method is similar to that of the original Lesk algorithm. The semantic relatedness of each sense of the target word is measured relative to every noun sense of the words in the window of context. The scores associated with each combination of senses are summed and used to assign a value to each candidate sense for the target word. The sense with the highest score is then assigned to the target word.

Candidate senses are scored by the original Adapted Lesk Algorithm using gloss overlaps. However, any measure of semantic relatedness could be used since the gloss overlaps simply produce a numeric score that indicates how many overlaps there are between the glosses of the senses of the target word and the glosses of the senses of the words in the window of context. The larger these scores, the more related the words. In this paper we extend the Adapted Lesk framework such that we can plug other measures of relatedness into the algorithm in place of gloss overlaps. All of the code used to calculate these measures and perform word sense disambiguation is available from the author's web pages.

6 Experimental Data

We compare the different measures of relatedness by employing them in the Adapted Lesk framework in place of gloss overlaps. We carried out word sense disambiguation experiments using the noun data from the English lexical sample task of SENSEVAL-2. In particular, this consisted of of 1,754 *instances* from the evaluation portion of this data, where each instance is made up of three to four sentences where a single word (the target word) has been manually assigned its most appropriate sense from WordNet.

This results in 29 nouns that serve as target words, each of which has between 1 and 14 possible WordNet senses. Table 1 lists the base form of the target words and the number of instances available for that word. It also specifies the number of WordNet senses for the base form of the target word in column *WN* and the the total number of candidate senses considered for each word in column *cand*. The number of candidate senses is greater than the number of possible senses of the base form because target words may appear in multiple forms over a set of instances.

These various forms consist of morphological variants and compounds. For example, while *art* is the base form of one of the target words, it also occurs as a target word in the data as *arts* and *art gallery*. WordNet has a separate sense inventory for each form, so we will consider additional or different sets of senses depending on the form of the target word. The number of candidate senses shown is the total number of candidate senses considered across all the instances and is not specific to any particular instance or form of the word.

7 Experiments and Results

Our empirical evaluation follows the model of Budanitsky and Hirst [3], who compare the five measures of relatedness previously described when applied to context sensitive spelling correction. We use those same measures in this study in addition to the gloss overlaps from the Adapted Lesk Algorithm.

We use the local approach from Adapted Lesk with a three word window of context. The framework of Adapted Lesk was generalized so that it could perform word sense disambiguation based on each of the semantic relatedness measures discussed above in addition to the original gloss overlap measure.

Results are reported as per word and overall accuracy, where the number of correct instances is divided by the total number of instances. Table 1 shows the disambiguation accuracy attained when using each of the different measures of semantic relatedness. The results for the Resnik, Jiang–Conrath, Lin, Leacock–Chodorow, Hirst–St.Onge, and gloss overlap measures are shown in columns *Res*, *Jcn*, *Lin*, *Lch*, *Hso* and *Lesk*, respectively. The highest accuracy achieved for each word is shown in bold face, while the least is italicized. The overall accuracy for each measure is shown at the bottom of the table.

The measures of Resnik, Jiang–Conrath and Lin depend upon the corpus used to estimate information content. We carried out experiments using five sources of information content: SemCor (with and without sense tags), the Brown Corpus [5], the Penn [11], and the British National Corpus.

The Brown Corpus is a 1,000,000 token corpus of balanced English. SemCor is a sense tagged subset of the Brown Corpus that consists of about 200,000 sense–tagged tokens, many of which are associated with concepts that only occur one time. The Penn Treebank is a 1,000,000 token corpus of English taken from the Wall Street Journal. The British National Corpus (BNC) is by far the largest of these corpora as it is a 100,000,000 token balanced sample of English.

Table 1. Experimental Results

Word	Instance	Senses		Measures of Relatedness					
	count	WN	cand	Res	Jcn	Lin	Lch	Hso	Lesk
art	98	4	14	0.41	0.54	0.42	0.44	*0.40*	**0.61**
authority	92	7	9	*0.14*	0.16	0.17	0.19	0.20	**0.27**
bar	151	13	21	0.21	0.23	**0.25**	*0.18*	**0.25**	0.21
bum	44	4	4	0.20	**0.73**	0.41	0.59	0.31	*0.13*
chair	69	4	7	0.37	0.33	0.46	*0.21*	0.44	**0.84**
channel	73	7	13	0.15	0.15	0.16	**0.23**	0.20	*0.10*
child	63	4	5	0.27	0.43	0.38	*0.02*	0.16	**0.62**
church	64	3	9	*0.37*	0.41	*0.37*	0.41	**0.48**	0.38
circuit	85	6	15	0.43	0.51	0.48	*0.34*	0.41	**0.53**
day	134	10	18	*0.12*	**0.43**	0.32	0.28	0.19	0.15
detention	32	2	5	0.61	0.81	0.61	*0.52*	0.63	**0.88**
dyke	28	2	2	0.73	0.86	0.77	*0.46*	0.61	**0.89**
facility	58	5	7	0.24	**0.34**	0.29	*0.21*	0.23	0.29
fatigue	43	4	6	*0.16*	0.42	0.22	**0.77**	0.44	**0.77**
feeling	51	6	7	*0.22*	**0.55**	0.27	0.53	0.26	0.49
grip	42	7	8	**0.22**	0.19	0.19	0.17	0.18	*0.12*
hearth	32	3	3	0.43	0.72	0.59	*0.38*	0.42	**0.75**
holiday	31	2	3	**0.55**	*0.16*	0.32	**0.55**	**0.55**	*0.16*
lady	53	3	8	0.36	*0.17*	0.19	**0.42**	0.36	*0.17*
material	69	5	10	0.44	**0.55**	0.44	0.40	0.38	*0.29*
mouth	57	8	8	0.12	0.12	0.11	*0.05*	0.20	**0.46**
nation	37	4	6	*0.18*	0.35	0.26	0.22	0.26	**0.59**
nature	44	5	6	0.10	0.11	*0.05*	0.11	**0.18**	0.16
post	78	8	12	0.16	**0.35**	0.19	*0.09*	0.15	0.31
restraint	45	6	7	0.31	**0.40**	0.33	0.36	0.30	*0.16*
sense	50	5	11	0.49	*0.40*	0.51	*0.40*	0.43	**0.50**
spade	33	3	4	**0.70**	*0.15*	0.56	0.21	0.40	0.59
stress	39	5	5	0.32	0.38	0.33	**0.44**	0.39	*0.31*
yew	28	2	3	0.66	0.79	0.73	*0.57*	0.70	**0.86**
Total	1723			*0.295*	0.380	0.331	0.305	0.316	**0.391**

We report the best results for each of these three measures in Table 1. The results for Jiang–Conrath and Lin are based on estimates of information content from the British National Corpus, while those of Resnik are based on the sense tagged version of SemCor. In addition, we provide the overall accuracies attained by each of these measures when their information content is determined from each of the indicated corpora in Table 2. Leacock–Chodorow, Hirst–St. Onge, and Adapted Lesk are not included here since they do not employ corpus information but instead only depend on information found in WordNet.

Table 2. Overall Accuracy Using Different Sources of Information Content

Corpus	Measure of Relatedness		
	Res	Jcn	Lin
SemCor	**0.295**	0.330	0.328
SemCor (untagged)	**0.295**	0.330	0.320
Brown Corpus	0.290	0.363	**0.331**
Penn Treebank	0.292	**0.380**	0.329
BNC	0.290	**0.380**	**0.331**

8 Analysis and Discussion

The gloss overlaps of Adapted Lesk and the Jiang–Conrath measure result in disambiguation that was significantly more accurate than the rest of the measures. The gloss overlaps of Adapted Lesk result in the highest overall accuracy (.391). In addition, it is the most accurate method for 13 of the 29 words. The next most accurate method overall was that of Jiang–Conrath when information content was estimated from the Penn Treebank or BNC (.380). This proved to be the most accurate method for 7 individual words.

Of the three measures based on information content, Jiang–Conrath was the only one that showed significant variations based on the corpora from which information content was estimated. It was most accurate with the British National Corpus and the Penn Treebank, and least accurate with SemCor (tagged and untagged).

Despite the very close similarity between the formulation of Jiang–Conrath and the Lin measure, there was a significant difference between them in accuracy for all corpora except SemCor (tagged and untagged). The accuracy of the Lin measure did not vary much with information content estimated from different corpora. The highest accuracy it attained was .331 and the lowest was .320. It was the least extreme of all the measures in that it was most accurate for one word and was the least accurate method for only two words. The Hirst–St. Onge measure (.316) was similarly conservative in that it was most accurate for four words and least accurate for just one.

The Leacock–Chodorow measure (.305) was most accurate for five words but least accurate for twelve words. Its overall accuracy was slightly higher than that of Resnik (.295), which was most accurate for three words and least accurate for six. Like the Lin measure, the accuracy associated with the Resnik measure did not vary a great deal with different sources of estimates for information content. The highest level was .295 and the lowest was .290.

These measures use a variety of different sources of information to determine the semantic relatedness of words. Leacock–Chodorow and Hirst–St. Onge rely on the structure of concept hierarchies; Resnik, Jiang–Conrath, and Lin augment this concept hierarchy with information content values estimated from corpora; and the Adapted Lesk Algorithm relies on gloss overlaps from WordNet.

The results attained by Leacock–Chodorow and Resnik suggest that simply relying on the concept hierarchy structure or information content values is not sufficient. The Jiang–Conrath measure combines the structure of WordNet with information content values taken from corpora and does extremely well and outperforms all other measures except the gloss overlaps of Adapted Lesk. This is an interesting result since the gloss overlaps are a completely different source of information.

8.1 Information Content Variations

We estimated information content from a number of corpora in order to study the effect of different amounts and types of data on disambiguation accuracy. Resnik's original experiments were with the Brown Corpus, while Lin and Jiang–Conrath used the sense–tagged version of SemCor.

We wanted to determine the effect of sense–tagged text on information content based measures. We expected sense–tagged text to be the best source of information content values, since each sense–tag represents a single concept and estimates derived from such data should be very reliable. We carried out disambiguation experiments using information content derived from the sense–tagged version of SemCor and then we repeated the experiment after removing the sense–tags.

Curiously enough, none of the three information content based measures performed significantly differently with the tagged and untagged versions of SemCor. We believe this is because many of the sense–tags in SemCor occur only once, thus creating a fairly sparse source of data. As a result the information content of the sense–tagged corpora is not significantly different than that of the untagged version.

We also wanted to assess the impact of increasing the size of the corpus from which information content values are estimated. Our initial experiments were done with SemCor, which has about 200,000 tokens. When carrying out the same experiments using the 1,000,000 token Brown Corpus and Penn Treebank, we observed that only Jiang–Conrath showed an increase in accuracy. Curiously enough, it performed considerably better with the Penn Treebank (.380) than it did with the more balanced Brown Corpus (.363). We say this is curious since

the SENSEVAL-2 data does not seem terribly similar to typical Penn Treebank text.

Both Resnik and Lin performed at nearly the same level of accuracy regardless of the corpora from which the information content values were estimated. The Resnik measure was most accurate (.295) with SemCor whether it was sense–tagged or not, and least accurate with the Brown Corpus and BNC (.290). The Lin Measure was least accurate with the untagged version of SemCor (.320) and most accurate with the Brown Corpus and BNC (.331).

Our final experiment was with the British National Corpus (BNC), which is a 100,000,000 token sample of English. Despite the huge increase in size, Jiang–Conrath performed at the same level of accuracy as achieved with the Penn Treebank, and Resnik and Lin attained accuracy equal to (or less than) that of SemCor. Thus, the very large increase in size of the corpus did not yield any benefit for word sense disambiguation. We are unclear as to why this would be the case, and consider this an important issue for future work.

8.2 Window Size Variations

The results reported earlier were based on window sizes of three, which include the target word and one content word to the right and left. We conducted several experiments with a window size of five, which includes the target word and two words to the right and left.

The most significant change in results when increasing the window size was with the Jiang–Conrath measure. For SemCor the accuracy rose to .341 with a window size of five (from .330) and for the BNC it attained .386 (from .380). As a result Jiang–Conrath achieves a level of accuracy that is essentially equal to that of the gloss overlaps of Adapted Lesk. The change of window size from three to five did not change the accuracy of Adapted Lesk.

The Lin measure with information content estimated from the Brown Corpus improved, rising from .331 to .341. The Resnik measure improved with respect to the Penn Treebank, which achieved .302 with a window size of five versus .292 with a window of three. In addition, the accuracy of the Hirst–St. Onge improved to .333 when the window size is five.

Thus, it appears that among the information content measures of Resnik, Lin, and Jiang–Conrath, the latter is the most able to take advantage of increased amounts of information. It is most accurate with a window size of five where its information content is estimated from the British National Corpus (.386). Given this same combination of window size and corpus, Lin achieves accuracy of .334 and Resnik reaches .298.

9 Related Work

A number of other methods to measure semantic relatedness of words have been proposed and used for word sense disambiguation. Agirre and Rigau [1] do not exactly describe a measure of semantic relatedness. Rather they introduce a

notion of *conceptual density* and use this in the process of word sense disambiguation. This notion of *conceptual density* is again based on the WordNet *is-a* hierarchy. The process of disambiguation of a target word starts by considering all the possible senses of the target word and the senses of the words in the window of context of the target word (window sizes of five to 30 words were considered in their experiments). Sub-hierarchies in the WordNet *is-a* taxonomy are then determined, such that each sub-hierarchy contains one of the senses of the target word along with senses of word in the context. The conceptual density for each of these sub-hierarchies is computed as the ratio of the average number of hypernyms per node for the senses of the context words to the average number of hypernyms per node for all nodes of the sub-hierarchy. This ratio gives us the distribution or density of these senses in the sub–hierarchy. The sense of the target word in the sub–hierarchy with the largest conceptual density is selected as the implied sense. Though this method does not specify an exact formula for semantic relatedness of words, it appears to be built upon node counting techniques for measuring semantic relatedness and gives us yet another way to use the WordNet *is–a* hierarchy for word sense disambiguation.

Leacock and Chodorow [8] have used their measure to augment a supervised approach to word sense disambiguation that relies on local context, which are features that occur in close proximity to the target word. They use their measure (as well as Resnik's) to determine the relatedness between a noun in each test instance with nouns in the training data. If there is a noun in a test instance that does not occur in the training data, then the most related noun found in the training data is substituted in order to allow for disambiguation to proceed.

Lin [10] also used his measure of semantic relatedness to perform the task of word sense disambiguation. However, unlike the procedure followed in this paper, he used his measure of semantic relatedness to generate a list of local contexts for each target word. This list of context would then restrict the possibility of what could appear in the context of a given word for a particular sense. This was used to disambiguate new instances of the word.

10 Future Work

The two most accurate methods in this study were quite dissimilar. Adapted Lesk gloss overlaps are based on the definitions found within WordNet, while the measure of Jiang–Conrath is based on the concept hierarchy of WordNet and corpus statistics. This suggests that some combination of gloss overlaps, information content, and path lengths might result in improved accuracy.

We are aware that our method of estimating information content employs a different counting scheme than described by Resnik. In short, we do not divide frequency counts of word types by the number of associated concepts while Resnik does. We will carry out these same experiments using Resnik's estimation scheme. Our expectation is that the results will not vary significantly, since in general we believe that the concept counts are fairly noisy regardless of how they are made.

One of the curious results of these experiments was how little disambiguation accuracy was affected by changing the corpora from which information content values are estimated. The Resnik and Lin measures were fairly static in their performance regardless of the source of these estimates. The Jiang–Conrath measure improved with increasing corpus size, except when increasing to the very large British National Corpus, where it resulted in the same disambiguation accuracy as information content arrived at from the 100 times smaller Penn Treebank. Our next step it to estimate information content from corpora that is more like the data we are disambiguating to see if this changes the results.

A related point concerns the relatively low impact achieved by increasing the window size. We are curious as to why such a large increase in the information available to the disambiguation process would result in such minimal improvements. One possibility is that the very immediate context provides overwhelming evidence that is difficult to improve upon. In order to evaluate this idea we will carry out experiments using a two word window, that is the target word and one content word that precedes it.

The difficulties of using WordNet as a source of path lengths and gloss overlaps are well known. We have recently acquired Longman's Dictionary of Contemporary English (LDOCE) and intend to use its more limited concept hierarchy but richer and more regular glosses to carry out experiments similar to these. While the results of measures based on path lengths may suffer (since the hierarchy represented by LDOCE subject codes is fairly small) we are optimistic that the richer gloss information might result in better performance for Lesk inspired approaches.

11 Conclusions

We have shown that the Adapted Lesk Algorithm of Banerjee and Pedersen generalizes to a method of disambiguation based on semantic relatedness. We showed that several different measures of semantic relatedness work reasonably well in this framework, and that the gloss overlaps of Adapted Lesk and the Jiang–Conrath measure prove to be the most accurate for word sense disambiguation.

12 Acknowledgments

Satanjeev Banerjee is currently supported by the National Science Foundation under Grant No. REC–9979894. Ted Pedersen is partially supported by a National Science Foundation Faculty Early CAREER Development award (#0092784).

Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation or the official policies, either expressed or implied, of the sponsors or of the United States Government.

References

1. E. Agirre and G. Rigau. Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 16–22, Copenhagen, 1996.
2. S. Banerjee and T. Pedersen. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February 2002.
3. 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.
4. C. Fellbaum, editor. *WordNet: An electronic lexical database*. MIT Press, 1998.
5. W. Francis and H. Kucera. *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, 1982.
6. G. Hirst and D. St. Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. MIT Press, 1998.
7. 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.
8. C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 265–283. MIT Press, 1998.
9. 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.
10. 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.
11. M. Marcus, B. Santorini, and M. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.
12. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30, 1989.
13. 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.
14. P. Resnik. WordNet and class–based probabilities. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 239–263. MIT Press, 1998.