# Determining Word Sense Dominance Using a Thesaurus

**Saif Mohammad and Graeme Hirst**
Department of Computer Science
University of Toronto
Toronto, ON M5S 3G4, Canada
{smm,gh}@cs.toronto.edu

## Abstract

The degree of dominance of a sense of a word is the proportion of occurrences of that sense in text. We propose four new methods to accurately determine word sense dominance using raw text and a published thesaurus. Unlike the McCarthy et al. (2004) system, these methods can be used on relatively small target texts, without the need for a *similarly-sense-distributed* auxiliary text. We perform an extensive evaluation using artificially generated thesaurus-sense-tagged data. In the process, we create a word–category co-occurrence matrix, which can be used for unsupervised word sense disambiguation and estimating distributional similarity of word *senses*, as well.

## 1 Introduction

The occurrences of the senses of a word usually have skewed distribution in text. Further, the distribution varies in accordance with the domain or topic of discussion. For example, the 'assertion of illegality' sense of *charge* is more frequent in the judicial domain, while in the domain of economics, the 'expense/cost' sense occurs more often. Formally, the **degree of dominance of a** particular **sense** of a word (**target word**) in a given text (**target text**) may be defined as the ratio of the occurrences of the sense to the total occurrences of the target word. The sense with the highest dominance in the target text is called the **predominant sense** of the target word.

Determination of word sense dominance has many uses. An unsupervised system will benefit by backing off to the predominant sense in case of insufficient evidence. The dominance values may be used as prior probabilities for the different senses, obviating the need for labeled training data in a sense disambiguation task. Natural language systems can choose to ignore infrequent senses of words or consider only the most dominant senses (McCarthy et al., 2004). An unsupervised algorithm that discriminates instances into different usages can use word sense dominance to assign senses to the different clusters generated.

Sense dominance may be determined by simple counting in sense-tagged data. However, dominance varies with domain, and existing sense-tagged data is largely insufficient. McCarthy et al. (2004) automatically determine domain-specific predominant senses of words, where the domain may be specified in the form of an untagged target text or simply by name (for example, *financial* domain). The system (Figure 1) automatically generates a thesaurus (Lin, 1998) using a measure of distributional similarity and an untagged corpus. The target text is used for this purpose, provided it is large enough to learn a thesaurus from. Otherwise a large corpus with sense distribution similar to the target text (text pertaining to the specified domain) must be used.

The thesaurus has an entry for each word type, which lists a limited number of words (**neighbors**) that are distributionally most similar to it. Since Lin's distributional measure overestimates the distributional similarity of more-frequent word pairs (Mohammad and Hirst, Submitted), the neighbors of a word corresponding to the predominant sense are distributionally closer to it than those corresponding to any other sense. For each sense of a word, the distributional similarity scores of all its neighbors are summed using the semantic similarity of the word with the closest sense of the

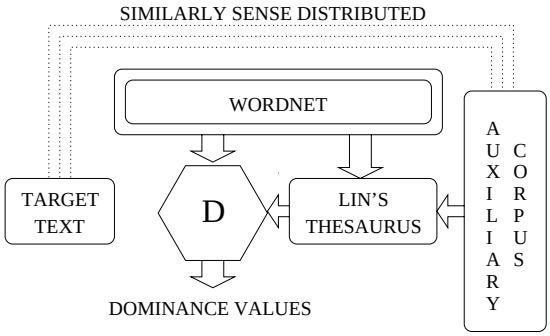


Figure 1: The McCarthy et al. system.

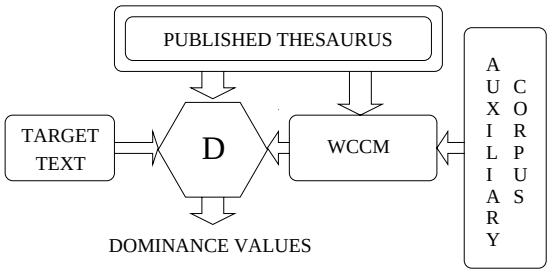


Figure 2: Our system.

neighbor as weight. The sense that gets the highest score is chosen as the predominant sense.

The McCarthy et al. system needs to re-train (create a new thesaurus) every time it is to determine predominant senses in data from a different domain. This requires large amounts of part-of-speech-tagged and chunked data from that domain. Further, the target text must be large enough to learn a thesaurus from (Lin (1998) used a 64-million-word corpus), or a large auxiliary text with a sense distribution similar to the target text must be provided (McCarthy et al. (2004) separately used 90-, 32.5-, and 9.1-million-word corpora).

By contrast, in this paper we present a method that accurately determines sense dominance even in relatively small amounts of target text (a few hundred sentences); although it does use a corpus, it does not require a *similarly-sense-distributed* corpus. Nor does our system (Figure 2) need any part-of-speech-tagged data (although that may improve results further), and it does not need to generate a thesaurus or execute any such time-intensive operation at run time. Our method stands on the hypothesis that words surrounding the target word are indicative of its intended sense, and that the dominance of a particular sense is proportional to the relative strength of association between it and co-occurring words in the target text. We therefore rely on first-order co-occurrences, which we believe are better indicators of a word's characteristics than second-order co-occurrences (distributionally similar words).

## 2 Thesauri

Published thesauri, such as *Roget's* and *Macquarie*, divide the English vocabulary into around a thousand **categories**. Each category has a list of semantically related words, which we will call **category terms** or **c-terms** for short. Words with multiple meanings may be listed in more than one category. For every word type in the vocabulary of the thesaurus, the index lists the categories that include it as a c-term. Categories roughly correspond to coarse senses of a word (Yarowsky, 1992), and the two terms will be used interchangeably. For example, in the *Macquarie Thesaurus*, *bark* is a c-term in the categories 'animal noises' and 'membrane'. These categories represent the coarse senses of *bark*. Note that published thesauri are structurally quite different from the "thesaurus" automatically generated by Lin (1998), wherein a word has exactly one entry, and its neighbors may be semantically related to it in any of its senses. All future mentions of *thesaurus* will refer to a published thesaurus.

While other sense inventories such as WordNet exist, use of a published thesaurus has three distinct advantages: (i) coarse senses—it is widely believed that the sense distinctions of WordNet are far too fine-grained (Agirre and Lopez de Lacalle Lekuona (2003) and citations therein); (ii) computational ease—with just around a thousand categories, the word–category matrix has a manageable size; (iii) widespread availability—thesauri are available (or can be created with relatively less effort) in numerous languages, while WordNet is available only for English and a few romance languages. We use the *Macquarie Thesaurus* (Bernard, 1986) for our experiments. It consists of 812 categories with around 176,000 c-terms and 98,000 word types. Note, however, that using a sense inventory other than WordNet will mean that we cannot directly compare performance with McCarthy et al. (2004), as that would require knowing exactly how thesaurus senses map to WordNet. Further, it has been argued that such a mapping across sense inventories is at best difficult and maybe impossible (Kilgarriff and Yallop (2001) and citations therein).

## 3 Co-occurrence Information

### 3.1 Word–Category Co-occurrence Matrix

The strength of association between a particular category of the target word and its co-occurring words can be very useful—calculating word sense dominance being just one application. To this end we create the **word–category co-occurrence matrix (WCCM)** in which one dimension is the list of all words $(w_1, w_2, \ldots)$ in the vocabulary, and the other dimension is a list of all categories $(c_1, c_2, \ldots)$.

| | $c_1$ | $c_2$ | $\ldots$ | $c_j$ | $\ldots$ |
|---|---|---|---|---|---|
| $w_1$ | $m_{11}$ | $m_{12}$ | $\ldots$ | $m_{1j}$ | $\ldots$ |
| $w_2$ | $m_{21}$ | $m_{22}$ | $\ldots$ | $m_{2j}$ | $\ldots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\ldots$ | $\ldots$ |
| $w_i$ | $m_{i1}$ | $m_{i2}$ | $\ldots$ | $m_{ij}$ | $\ldots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ |

A particular cell, $m_{ij}$, pertaining to word $w_i$ and category $c_j$, is the number of times $w_i$ occurs in a predetermined window around any c-term of $c_j$ in a text corpus. We will refer to this particular WCCM created after the *first* pass over the text as the **base WCCM**. A contingency table for any particular word $w$ and category $c$ (see below) can be easily generated from the WCCM by collapsing cells for all other words and categories into one and summing up their frequencies. The application of a suitable statistic will then yield the strength of association between the word and the category.

| | $c$ | $\neg c$ |
|---|---|---|
| $w$ | $n_{wc}$ | $n_{w\neg}$ |
| $\neg w$ | $n_{\neg c}$ | $n_{\neg\neg}$ |

Even though the base WCCM is created from unannotated text, and so is expected to be noisy, we argue that it captures strong associations reasonably accurately. This is because the errors in determining the true category that a word co-occurs with will be distributed thinly across a number of other categories (details in Section 3.2). Therefore, we can take a *second* pass over the corpus and determine the intended sense of each word using the word–category co-occurrence frequency (from the base WCCM) as evidence. We can thus create a newer, more accurate, **bootstrapped WCCM** by populating it just as mentioned earlier, except that this time counts of only the co-occurring word and the disambiguated category are incremented. The steps of word sense disambiguation and creating new bootstrapped WCCMs can be repeated until the bootstrapping fails to improve accuracy significantly.

The cells of the WCCM are populated using a large untagged corpus (usually different from the target text) which we will call the **auxiliary corpus**. In our experiments we use a subset (all except every twelfth sentence) of the *British National Corpus World Edition (BNC)* (Burnard, 2000) as the auxiliary corpus and a window size of $\pm 5$ words. The remaining one twelfth of the *BNC* is used for evaluation purposes. Note that if the target text belongs to a particular domain, then the creation of the WCCM from an auxiliary text of the same domain is expected to give better results than the use of a domain-free text.

### 3.2 Analysis of the Base WCCM

The use of untagged data for the creation of the base WCCM means that words that do not really co-occur with a certain category but rather do so with a homographic word used in a different sense will (erroneously) increment the counts corresponding to the category. Nevertheless, the strength of association, calculated from the base WCCM, of words that truly and strongly co-occur with a certain category will be reasonably accurate despite this noise.

We demonstrate this through an example. Assume that category $c$ has 100 c-terms and each c-term has 4 senses, only one of which corresponds to $c$ while the rest are randomly distributed among other categories. Further, let there be 5 sentences each in the auxiliary text corresponding to every c-term–sense pair. If the window size is the complete sentence, then words in 2,000 sentences will increment co-occurrence counts for $c$. Observe that 500 of these sentences truly correspond to category $c$, while the other 1500 pertain to about 300 other categories. Thus on average 5 sentences correspond to each category other than $c$. Therefore in the 2000 sentences, words that truly co-occur with $c$ will likely occur a large number of times, while the rest will be spread out thinly over 300 or so other categories.

We therefore claim that the application of a suitable statistic, such as odds ratio, will result in significantly large association values for word–category pairs where the word truly and strongly co-occurs with the category, and the effect of noise

will be insignificant. The word–category pairs having low strength of association will likely be adversely affected by the noise, since the amount of noise may be comparable to the actual strength of association. In most natural language applications, the strength of association is evidence for a particular proposition. In that case, even if association values from all pairs are used, evidence from less-reliable, low-strength pairs will contribute little to the final cumulative evidence, as compared to more-reliable, high-strength pairs. Thus even if the base WCCM is less accurate when generated from untagged text, it can still be used to provide association values suitable for most natural language applications. Experiments to be described in section 6 below substantiate this.

### 3.3 Measures of Association

The strength of association between a sense or category of the target word and its co-occurring words may be determined by applying a suitable statistic on the corresponding contingency table. Association values are calculated from observed frequencies ($n_{wc}, n_{\neg c}, n_{w\neg}$, and $n_{\neg\neg}$), marginal frequencies ($n_{w*} = n_{wc} + n_{w\neg}$; $n_{\neg *} = n_{\neg c} + n_{\neg\neg}$; $n_{*c} = n_{wc} + n_{\neg c}$; and $n_{*\neg} = n_{w\neg} + n_{\neg\neg}$), and the sample size ($N = n_{wc} + n_{\neg c} + n_{w\neg} + n_{\neg\neg}$). We provide experimental results using Dice coefficient (*Dice*), cosine (cos), pointwise mutual information (*pmi*), odds ratio (*odds*), Yule's coefficient of colligation (*Yule*), and phi coefficient ($\phi$)[1].

## 4 Word Sense Dominance

We examine each occurrence of the target word in a given untagged target text to determine dominance of any of its senses. For each occurrence $t'$ of a target word $t$, let $T'$ be the set of words (tokens) co-occurring within a predetermined window around $t'$; let $T$ be the union of all such $T'$ and let $\mathscr{X}_t$ be the set of all such $T'$. (Thus $|\mathscr{X}_t|$ is equal to the number of occurrences of $t$, and $|T|$ is equal to the total number of words (tokens) in the windows around occurrences of $t$.) We describe

[1] Measures of association (Sheskin, 2003):

$$\cos(w,c) = \frac{n_{wc}}{\sqrt{n_{w*}} \times \sqrt{n_{*c}}},\ pmi(w,c) = \log \frac{n_{wc} \times N}{n_{w*} \times n_{*c}},$$

$$odds(w,c) = \frac{n_{wc} \times n_{\neg\neg}}{n_{w\neg} \times n_{\neg c}},\ Yule(w,c) = \frac{\sqrt{odds(w,c)} - 1}{\sqrt{odds(w,c)} + 1},$$

$$Dice(w,c) = \frac{2 \times n_{wc}}{n_{w*} + n_{*c}},\ \phi(w,c) = \frac{(n_{wc} \times n_{\neg\neg}) - (n_{w\neg} \times n_{\neg c})}{\sqrt{n_{w*} \times n_{\neg *} \times n_{*c} \times n_{*\neg}}}$$

| | Weighted voting | Unweighted voting |
|---|---|---|
| Implicit sense disambiguation | $D_{I,W}$ | $D_{I,U}$ |
| Explicit sense disambiguation | $D_{E,W}$ | $D_{E,U}$ |

Figure 3: The four dominance methods.

four methods (Figure 3) to determine dominance ($D_{I,W}, D_{I,U}, D_{E,W}$, and $D_{E,U}$) and the underlying assumptions of each.

$D_{I,W}$ is based on the assumption that the more dominant a particular sense is, the greater the strength of its association with words that co-occur with it. For example, if most occurrences of *bank* in the target text correspond to 'river bank', then the strength of association of 'river bank' with all of *bank*'s co-occurring words will be larger than the sum for any other sense. Dominance $D_{I,W}$ of a sense or category ($c$) of the target word ($t$) is:

$$D_{I,W}(t,c) = \frac{\sum_{w \in T} A(w,c)}{\sum_{c' \in senses(t)} \sum_{w \in T} A(w,c')} \quad (1)$$

where $A$ is any one of the measures of association from section 3.3. Metaphorically, words that co-occur with the target word give a weighted vote to each of its senses. The weight is proportional to the strength of association between the sense and the co-occurring word. The dominance of a sense is the ratio of the total votes it gets to the sum of votes received by all the senses.

A slightly different assumption is that the more dominant a particular sense is, the greater the number of co-occurring words having highest strength of association with that sense (as opposed to any other). This leads to the following methodology. Each co-occurring word casts an equal, unweighted vote. It votes for that sense (and no other) of the target word with which it has the highest strength of association. The dominance $D_{I,U}$ of the sense is the ratio of the votes it gets to the total votes cast for the word (number of co-occurring words).

$$D_{I,U}(t,c) = \frac{|\{w \in T : Sns_1(w,t) = c\}|}{|T|} \quad (2)$$

$$Sns_1(w,t) = \underset{c' \in senses(t)}{\operatorname{argmax}}\ A(w,c') \quad (3)$$

Observe that in order to determine $D_{I,W}$ or $D_{I,U}$, we do not need to explicitly disambiguate

the senses of the target word's occurrences. We now describe alternative approaches that may be used for explicit sense disambiguation of the target word's occurrences and thereby determine sense dominance (the proportion of occurrences of that sense). $D_{E,W}$ relies on the hypothesis that the intended sense of any occurrence of the target word has highest strength of association with its co-occurring words.

$$D_{E,W}(t,c) = \frac{|\{T' \in \mathscr{X}_t : Sns_2(T',t) = c\}|}{|\mathscr{X}_t|} \quad (4)$$

$$Sns_2(T',t) = \underset{c' \in senses(t)}{\operatorname{argmax}} \sum_{w \in T'} A(w,c') \quad (5)$$

Metaphorically, words that co-occur with the target word give a weighted vote to each of its senses just as in $D_{I,W}$. However, votes from co-occurring words in an occurrence are summed to determine the intended sense (sense with the most votes) of the target word. The process is repeated for all occurrences that have the target word. If each word that co-occurs with the target word votes as described for $D_{I,U}$, then the following hypothesis forms the basis of $D_{E,U}$: in a particular occurrence, the sense that gets the maximum votes from its neighbors is the intended sense.

$$D_{E,U}(t,c) = \frac{|\{T' \in \mathscr{X}_t : Sns_3(T',t) = c\}|}{|\mathscr{X}_t|} \quad (6)$$

$$Sns_3(T',t) = \underset{c' \in senses(t)}{\operatorname{argmax}} |\{w \in T' : Sns_1(w,t) = c'\}| \quad (7)$$

In methods $D_{E,W}$ and $D_{E,U}$, the dominance of a sense is the proportion of occurrences of that sense.

The degree of dominance provided by all four methods has the following properties: (i) The dominance values are in the range 0 to 1—a score of 0 implies lowest possible dominance, while a score of 1 means that the dominance is highest. (ii) The dominance values for all the senses of a word sum to 1.

## 5 Pseudo-Thesaurus-Sense-Tagged Data

To evaluate the four dominance methods we would ideally like sentences with target words annotated with senses from the thesaurus. Since human annotation is both expensive and time intensive, we present an alternative approach of artificially generating thesaurus-sense-tagged data following the ideas of Leacock et al. (1998). Around 63,700 of the 98,000 word types in the *Macquarie Thesaurus* are **monosemous**—listed under just one of the 812 categories. This means that on average around 77 c-terms per category are monosemous. **Pseudo-thesaurus-sense-tagged (PTST) data** for a non-monosemous target word $t$ (for example, *brilliant*) used in a particular sense or category $c$ of the thesaurus (for example, 'intelligence') may be generated as follows. Identify monosemous c-terms (for example, *clever*) belonging to the same category as $c$. Pick sentences containing the monosemous c-terms from an untagged auxiliary text corpus.

*Hermione had a **clever** plan.*

In each such sentence, replace the monosemous word with the target word $t$. In theory the c-terms in a thesaurus are near-synonyms or at least strongly related words, making the replacement of one by another acceptable. For the sentence above, we replace *clever* with *brilliant*. This results in (artificial) sentences with the target word used in a sense corresponding to the desired category. Clearly, many of these sentences will not be linguistically well formed, but the non-monosemous c-term used in a particular sense is likely to have similar co-occurring words as the monosemous c-term of the same category.[2] This justifies the use of these pseudo-thesaurus-sense-tagged data for the purpose of evaluation.

We generated PTST test data for the head words in SENSEVAL-1 English lexical sample space[3] using the *Macquarie Thesaurus* and the held out subset of the *BNC* (every twelfth sentence).

## 6 Experiments

We evaluate the four dominance methods, like McCarthy et al. (2004), through the accuracy of a naive sense disambiguation system that always gives out the predominant sense of the target word. In our experiments, the predominant sense is determined by each of the four dominance methods, individually. We used the following setup to study the effect of sense distribution on performance.

[2] Strong collocations are an exception to this, and their effect must be countered by considering larger window sizes. Therefore, we do not use a window size of just one or two words on either side of the target word, but rather windows of $\pm 5$ words in our experiments.

[3] SENSEVAL-1 head words have a wide range of possible senses, and availability of alternative sense-tagged data may be exploited in the future.

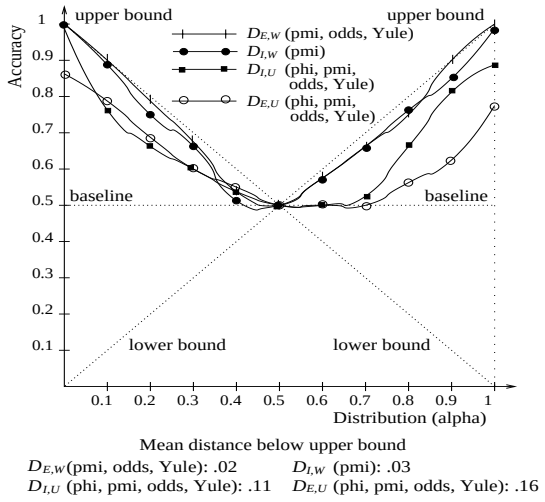


Figure 4: Best results: four dominance methods

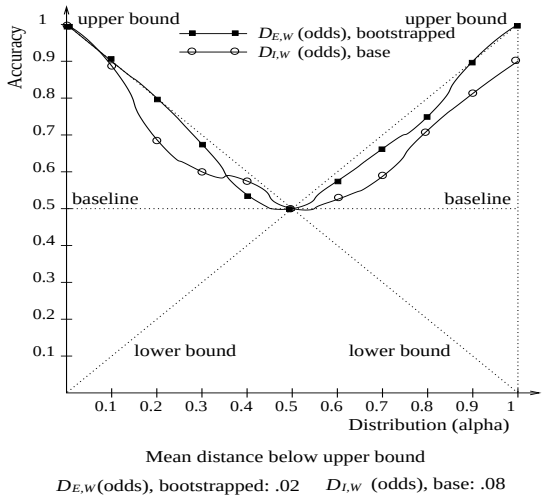


Figure 5: Best results: base vs. bootstrapped

### 6.1 Setup

For each target word for which we have PTST data, the two most dominant senses are identified, say $s_1$ and $s_2$. If the number of sentences annotated with $s_1$ and $s_2$ is $x$ and $y$, respectively, where $x > y$, then all $y$ sentences of $s_2$ and the first $y$ sentences of $s_1$ are placed in a **data bin**. Eventually the bin contains an equal number of PTST sentences for the two most dominant senses of each target word. Our data bin contained 17,446 sentences for 27 nouns, verbs, and adjectives. We then generate different test data sets $d_\alpha$ from the bin, where $\alpha$ takes values $0, .1, .2, \ldots, 1$, such that the fraction of sentences annotated with $s_1$ is $\alpha$ and those with $s_2$ is $1-\alpha$. Thus the data sets have different dominance values even though they have the same number of sentences—half as many in the bin.

Each data set $d_\alpha$ is given as input to the naive sense disambiguation system. If the predominant sense is correctly identified for all target words, then the system will achieve highest accuracy, whereas if it is falsely determined for all target words, then the system achieves the lowest accuracy. The value of $\alpha$ determines this **upper bound** and **lower bound**. If $\alpha$ is close to 0.5, then even if the system correctly identifies the predominant sense, the naive disambiguation system cannot achieve accuracies much higher than 50%. On the other hand, if $\alpha$ is close to 0 or 1, then the system may achieve accuracies close to 100%. A disambiguation system that randomly chooses one of the two possible senses for each occurrence of the target word will act as the baseline. Note that no matter what the distribution of the two senses ($\alpha$), this system will get an accuracy of 50%.

### 6.2 Results

Highest accuracies achieved using the four dominance methods and the measures of association that worked best with each are shown in Figure 4. The table below the figure shows **mean distance below upper bound (MDUB)** for all $\alpha$ values considered. Measures that perform almost identically are grouped together and the MDUB values listed are averages. The window size used was $\pm 5$ words around the target word. Each dataset $d_\alpha$, which corresponds to a different target text in Figure 2, was processed in less than 1 second on a 1.3GHz machine with 16GB memory. Weighted voting methods, $D_{E,W}$ and $D_{I,W}$, perform best with MDUBs of just .02 and .03, respectively. Yule's coefficient, odds ratio, and pmi give near-identical, maximal accuracies for all four methods with a slightly greater divergence in $D_{I,W}$, where pmi does best. The $\phi$ coefficient performs best for unweighted methods. Dice and cosine do only slightly better than the baseline. In general, results from the method–measure combinations are symmetric across $\alpha = 0.5$, as they should be.

Marked improvements in accuracy were achieved as a result of bootstrapping the WCCM (Figure 5). Most of the gain was provided by the first iteration itself, whereas further iterations resulted in just marginal improvements. All bootstrapped results reported in this paper pertain to just one iteration. Also, the bootstrapped WCCM is 72% smaller, and 5 times faster at processing the data sets, than the base WCCM, which has many non-zero cells even though the corresponding word and category never actually co-occurred (as mentioned in Section 3.2 earlier).

### 6.3 Discussion

Considering that this is a completely unsupervised approach, not only are the accuracies achieved using the weighted methods well above the baseline, but also remarkably close to the upper bound. This is especially true for $\alpha$ values close to 0 and 1. The lower accuracies for $\alpha$ near 0.5 are understandable as the amount of evidence towards both senses of the target word are nearly equal.

Odds, pmi, and Yule perform almost equally well for all methods. Since the number of times two words co-occur is usually much less than the number of times they occur individually, pmi tends to approximate the logarithm of odds ratio. Also, Yule is a derivative of odds. Thus all three measures will perform similarly in case the co-occurring words give an unweighted vote for the most appropriate sense of the target as in $D_{I,U}$ and $D_{E,U}$. For the weighted voting schemes, $D_{I,W}$ and $D_{E,W}$, the effect of scale change is slightly higher in $D_{I,W}$ as the weighted votes are summed over the complete text to determine dominance. In $D_{E,W}$ the small number of weighted votes summed to determine the sense of the target word may be the reason why performances using pmi, Yule, and odds do not differ markedly. Dice coefficient and cosine gave below-baseline accuracies for a number of sense distributions. This suggests that the normalization[4] to take into account the frequency of individual events inherent in the Dice and cosine measures may not be suitable for this task.

The accuracies of the dominance methods remain the same if the target text is partitioned as per the target word, and each of the pieces is given individually to the disambiguation system. The average number of sentences per target word in each dataset $d_\alpha$ is 323. Thus the results shown above correspond to an average target text size of only 323 sentences.

We repeated the experiments on the base WCCM after filtering out (setting to 0) cells with frequency less than 5 to investigate the effect on accuracies and gain in computation time (proportional to size of WCCM). There were no marked changes in accuracy but a 75% reduction in size of the WCCM. Using a window equal to the complete sentence as opposed to $\pm 5$ words on either side of the target resulted in a drop of accuracies.

[4] If two events occur individually a large number of times, then they must occur together much more often to get substantial association scores through pmi or odds, as compared to cosine or the Dice coefficient.

## 7 Related Work

The WCCM has similarities with latent semantic analysis, or LSA, and specifically with work by Schütze and Pedersen (1997), wherein the dimensionality of a word–word co-occurrence matrix is reduced to create a word–concept matrix. However, there is no non-heuristic way to determine when the dimension reduction should stop. Further, the generic concepts represented by the reduced dimensions are not interpretable, i.e., one cannot determine which concepts they represent in a given sense inventory. This means that LSA cannot be used directly for tasks such as unsupervised sense disambiguation or determining semantic similarity of known concepts. Our approach does not have these limitations.

Yarowsky (1992) uses the product of a mutual information–like measure and frequency to identify words that best represent each category in the *Roget's Thesaurus* and uses these words for sense disambiguation with a Bayesian model. We improved the accuracy of the WCCM using simple bootstrapping techniques, used all the words that co-occur with a category, and proposed four new methods to determine sense dominance—two of which do explicit sense disambiguation. Véronis (2005) presents a graph theory–based approach to identify the various senses of a word in a text corpus without the use of a dictionary. Highly interconnected components of the graph represent the different senses of the target word. The node (word) with the most connections in a component is representative of that sense and its associations with words that occur in a test instance are used as evidence for that sense. However, these associations are at best only rough estimates of the associations between the sense and co-occurring words, since a sense in his system is represented by a single (possibly ambiguous) word. Pantel (2005) proposes a framework for ontologizing lexical resources. For example, co-occurrence vectors for the nodes in WordNet can be created using the co-occurrence vectors for words (or lexicals). However, if a leaf node has a single lexical, then once the appropriate co-occurring words for this node are identified (coup phase), they are assigned the same co-occurrence counts as that of the lexical.[5]

[5] A word may have different, stronger-than-chance strengths of association with multiple *senses* of a lexical. These are different from the association of the word with the *lexical*.

## 8 Conclusions and Future Directions

We proposed a new method for creating a word–category co-occurrence matrix (WCCM) using a published thesaurus and raw text, and applying simple sense disambiguation and bootstrapping techniques. We presented four methods to determine degree of dominance of a sense of a word using the WCCM. We automatically generated sentences with a target word annotated with senses from the published thesaurus, which we used to perform an extensive evaluation of the dominance methods. We achieved near-upper-bound results using all combinations of the the weighted methods ($D_{I,W}$ and $D_{E,W}$) and three measures of association (odds, pmi, and Yule).

We cannot compare accuracies with McCarthy et al. (2004) because use of a thesaurus instead of WordNet means that knowledge of exactly how the thesaurus senses map to WordNet is required. We used a thesaurus as such a resource, unlike WordNet, is available in more languages, provides us with coarse senses, and leads to a smaller WCCM (making computationally intensive operations viable). Further, unlike the McCarthy et al. system, we showed that our system gives accurate results without the need for a large *similarly-sense-distributed* text or retraining. The target texts used were much smaller (few hundred sentences) than those needed for automatic creation of a thesaurus (few million words).

The WCCM has a number of other applications, as well. The strength of association between a word and a word sense can be used to determine the (more intuitive) distributional similarity of *word senses* (as opposed to *words*). Conditional probabilities of lexical features can be calculated from the WCCM, which in turn can be used in unsupervised sense disambiguation. In conclusion, we provided a framework for capturing distributional properties of word *senses* from raw text and demonstrated one of its uses—determining word sense dominance.

### Acknowledgments

We thank Diana McCarthy, Afsaneh Fazly, and Suzanne Stevenson for their valuable feedback. This research is financially supported by the Natural Sciences and Engineering Research Council of Canada and the University of Toronto.

## References

Eneko Agirre and O. Lopez de Lacalle Lekuona. 2003. Clustering WordNet word senses. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'03)*, Bulgaria.

J.R.L. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.

Lou Burnard. 2000. *Reference Guide for the British National Corpus (World Edition)*. Oxford University Computing Services.

Adam Kilgarriff and Colin Yallop. 2001. What's in a thesaurus. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pages 1371–1379, Athens, Greece.

Claudia Leacock, Martin Chodrow, and George A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-98)*, pages 768–773, Montreal, Canada.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 280–267, Barcelona, Spain.

Saif Mohammad and Graeme Hirst. Submitted. Distributional measures as proxies for semantic relatedness.

Patrick Pantel. 2005. Inducing ontological co-occurrence vectors. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, pages 125–132, Ann Arbor, Michigan.

Hinrich Schütze and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retreival. *Information Processing and Management*, 33(3):307–318.

David Sheskin. 2003. *The handbook of parametric and nonparametric statistical procedures*. CRC Press, Boca Raton, Florida.

Jean Véronis. 2005. Hyperlex: Lexical cartography for information retrieval. *To appear in Computer Speech and Language. Special Issue on Word Sense Disambiguation*.

David Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, pages 454–460, Nantes, France.