

# Estimating Semantic Distance Using Soft Semantic Constraints in Knowledge-Source–Corpus Hybrid Models

Yuval Marton<sup>\*†</sup>, Saif Mohammad<sup>†</sup>, and Philip Resnik<sup>\*†</sup>

<sup>\*</sup>Department of Linguistics and

<sup>†</sup>Laboratory for Computational Linguistics and Information Processing,  
Institute for Advanced Computer Studies.

University of Maryland, College Park, MD 20742-7505, USA.

{ymarton, saif, resnik}@umiacs.umd.edu

## Abstract

Strictly corpus-based measures of semantic distance conflate co-occurrence information pertaining to the many possible senses of target words. We propose a corpus–thesaurus hybrid method that uses soft constraints to generate word-sense-aware distributional profiles (DPs) from coarser “concept DPs” (derived from a Roget-like thesaurus) and sense-unaware traditional word DPs (derived from raw text). Although it uses a knowledge source, the method is not vocabulary-limited: if the target word is not in the thesaurus, the method falls back gracefully on the word’s co-occurrence information. This allows the method to access valuable information encoded in a lexical resource, such as a thesaurus, while still being able to effectively handle domain-specific terms and named entities. Experiments on word-pair ranking by semantic distance show the new hybrid method to be superior to others.

## 1 Introduction

Semantic distance is a measure of the closeness in meaning of two concepts. People are consistent judges of semantic distance. For example, we can easily tell that the concepts of “exercise” and “jog” are closer in meaning than “exercise” and “theater”. Studies asking native speakers of a language to rank word pairs in order of semantic distance confirm this—average inter-annotator correlation on ranking word pairs in order of semantic distance has been repeatedly shown to be around 0.9 (Rubenstein and Goodenough, 1965; Resnik, 1999).

A number of natural language tasks such as machine translation (Lopez, 2008) and word sense disambiguation (Banerjee and Pedersen, 2003; McCarthy, 2006), can be framed as semantic distance problems. Thus, developing automatic measures that are in-line with human notions of semantic distance has received much attention. These automatic approaches to semantic distance rely on manually created lexical resources such as WordNet, large amounts of text corpora, or both.

WordNet-based information content measures have been successful (Hirst and Budanitsky, 2005), but there are significant limitations on their applicability. They can be applied only if a WordNet exists for the language of interest (which is not the case for the “low-density” languages); and even if there is a WordNet, a number of domain-specific terms may not be encoded in it. On the other hand, corpus-based distributional measures of semantic distance, such as cosine and  $\alpha$ -skew divergence (Dagan et al., 1999), rely on raw text alone (Weeds et al., 2004; Mohammad, 2008). However, when used to rank word pairs in order of semantic distance or correct real-word spelling errors, they have been shown to perform poorly (Weeds et al., 2004; Mohammad and Hirst, 2006).

Mohammad and Hirst (2006) and Patwardhan and Pedersen (2006) argued that word sense ambiguity is a key reason for the poor performance of traditional distributional measures, and they proposed hybrid approaches that are distributional in nature, but also make use of information in lexical resources such as published thesauri and WordNet. However, both these approaches can be applied to estimate the semantic distance between two terms *only* if both terms exist in the lexical resource they rely on. We know lexical resources tend to have limited vocabulary and a large number of domain-

specific terms are usually not included.

It should also be noted that similarity values from different distance measures are not comparable (even after normalization to the same scale), that is, a similarity score of .75 as per one distance measure does not correspond to the same semantic distance as a similarity score of .75 from another distance measure.<sup>1</sup> Thus if one uses two independent distance measures, in this case: one resource-reliant and one only corpus-dependent, then these two measures are not comparable (and hence cannot be used in tandem), even if both rely—partially or entirely—on distributional corpus statistics.

We propose a hybrid semantic distance method that *inherently* combines the elements of a resource-reliant measure and a strictly corpus-dependent measure by imposing resource-reliant soft constraints on the corpus-dependent model. We choose the Mohammad and Hirst (2006) method as the resource-reliant method and not one of the WordNet-based measures because, unlike the WordNet-based measures, the Mohammad and Hirst method is distributional in nature and so lends itself immediately for combination with traditional distributional similarity measures. Our new hybrid method combines concept–word co-occurrence information (the Mohammad and Hirst distributional profiles of thesaurus concepts (DPC)) with word–word co-occurrence information, to generate word-sense-biased distributional profiles. The “pure” corpus-based distributional profile (a.k.a. *co-occurrence vector*, or *word association vector*), for some target word  $u$ , is biased with soft constraints towards each of the concepts  $c$  that list  $u$  in the thesaurus, to create a distributional profile that is specific to  $u$  in the sense that is most related to the other words listed under  $c$ .

Thus, this method can make more fine-grained distinctions than the Mohammad and Hirst method, and yet uses word sense information.<sup>2</sup> Our proposed method falls back gracefully to rely only on word-word co-occurrence information if any of the target terms is not listed in the lexical resource. Experiments on the word-pair ranking task

on three different datasets show that the our proposed hybrid measure outperforms all other comparable distance measures.

Mohammad and Hirst (2007) show that their method can be used to compute semantic distance in a resource poor language  $L_1$  by combining its text with a thesaurus in a resource-rich language  $L_2$  using an  $L_1$ – $L_2$  bilingual lexicon to create cross-lingual distributional profiles of concepts, that is,  $L_2$  word co-occurrence profiles of  $L_1$  thesaurus concepts. Since our method makes use of the Mohammad and Hirst DPCs, it can just as well make use of their cross-lingual DPCs, to compute semantic distance in a resource-poor language, just as they did. We leave that for future work.

## 2 Background and Related Work

Strictly speaking, semantic distance/closeness is a property of lexical units—a combination of the surface form and word sense.<sup>3</sup> Two terms are considered to be semantically close if there is a lexical semantic relation between them. Such a relation may be a classical relation such as hypernymy, troponymy, meronymy, and antonymy, or it may be what have been called an ad-hoc non-classical relation, such as cause-and-effect (Morris and Hirst, 2004). If the closeness in meaning is due to certain specific classical relations such as hypernymy and troponymy, then the terms are said to be semantically *similar*. Semantic *relatedness* is the term used to describe the more general form of semantic closeness caused by any semantic relation (Hirst and Budanitsky, 2005). So the nouns *liquid* and *water* are both semantically similar and semantically related, whereas the nouns *boat* and *rudder* are semantically related, but not similar.

The next three sub-sections describe three kinds of automatic distance measures: (1) lexical-resource-based measures that rely on a manually created resource such as WordNet; (2) corpus-based measures that rely only on co-occurrence statistics from large corpora; and (3) hybrid measures that are distributional in nature, and that also exploit the information in a lexical resource.

### 2.1 Lexical-resource-based measures

WordNet is a manually-created hierarchical network of nodes (taxonomy), where each node in

<sup>1</sup>All we can infer is that if  $w_1$  and  $w_2$  have a similarity score of .75 and  $w_3$  and  $w_4$  have a score of .5 by the *same* distance measure, then  $w_1$ – $w_2$  are closer in meaning than  $w_3$ – $w_4$ .

<sup>2</sup>Even though Mohammad and Hirst (2006) use thesaurus categories as coarse concepts, their algorithm can be applied using more finer-grained thesaurus word groupings (paragraphs and semicolon units), as well.

<sup>3</sup>The notion of semantic distance can be generalized, of course, to larger units such as phrases, sentences, passages, and so on (Landauer et al., 1998).

the network represents a fine-grained concept or word sense. An edge between two nodes represents a lexical semantic relation such as hypernymy and troponymy. WordNet-based measures consider two terms to be close if they occur close to each other in the network (connected by only a few arcs), if their definitions share many terms (Banerjee and Pedersen, 2003; Patwardhan and Pedersen, 2006), or if they share a lot of information (Lin, 1998; Resnik, 1999). The length of each arc/link (distance between nodes) can be assumed a unit length, or can be computed from corpus statistics. Within WordNet, the *is-a* hierarchy is much more well-developed than that of other lexical semantic relations. So, not surprisingly, the best WordNet-based measures are those that rely only on the *is-a* hierarchy. Therefore, they are good at measuring semantic similarity (e.g., *doctor-physician*), but not semantic relatedness (e.g., *doctor-scalpel*). Further, the measures can only be used in languages that have a (sufficiently developed) WordNet. WordNet sense information has been criticized to be too fine grained (Agirre and Lopez de Lacalle Lekuona, 2003; Navigli, 2006). See Hirst and Budanitsky (2005) for a comprehensive survey of WordNet-based measures.

## 2.2 Corpus-based measures

Strictly corpus-based measures of distributional similarity rely on the hypothesis that words that occur in similar context tend to be semantically close (Firth, 1957; Harris, 1940). The set of contexts of each target word  $u$  is represented by its distributional profile (DP)—the set of words that tend to co-occur with  $u$  within a certain distance, along with numeric scores signifying this co-occurrence tendency with  $u$ . Then measures such as cosine or  $\alpha$ -skew divergence are used to determine how close the DPs of the two target words are. See Section 3 for more details and related work. These measures are very appealing because they rely simply on raw text, but, as described earlier, when used to rank word pairs in order of semantic distance, or to correct real-word spelling errors, they perform poorly, compared to the WordNet-based measures. See Weeds et al. (2004), Mohammad (2008), and Curran (2004) for detailed surveys of distributional measures.

As Mohammad and Hirst (2006) point out, the DP of a word  $u$  conflates information about the potentially many senses of  $u$ . For example, con-

sider the following. The noun *bank* has two senses “river bank” and “financial institution”. Assume that *bank*, when used in the “financial institution” sense, co-occurred with the noun *money* 100 times in a corpus. Similarly, assume that *bank*, when used in the “river bank” sense, co-occurred with the noun *boat* 80 times. So the DP of *bank* will have co-occurrence information with *money* as well as *boat*:

DPW(*bank*):  
*money*,100; *boat*,80; *bond*,70; *fish*,77; ...

Assume that the DP of the word *ATM* is:

DPW(*ATM*):  
*money*,120; *boat*,0; *bond*,90; *fish*,0; ...

Thus the distributional distance of *bank* with *ATM* will be some sort of an average of the semantic distance between the “financial institution” and “ATM” senses and the semantic distance between the “river bank” and “ATM” senses. However, in various natural language tasks, we need the semantic distance between the intended senses of *bank* and *ATM*, which often also tends to be the semantic distance between their closest senses.

## 2.3 Hybrid measures

Both Mohammad and Hirst (2006) and Patwardhan and Pedersen (2006) proposed measures that are not only distributional in nature but also rely on a lexical resource to exploit the manually encoded information therein as well as to overcome the sense-conflation problem (described in section 2.2). Since we essentially combine the Mohammad and Hirst method with a “pure” word-based distributional measure to create our hybrid approach, we briefly describe their method here.

Mohammad and Hirst (2006) generate separate distributional profiles for the different senses of a word, without using any sense-annotated data. They use the categories in a Roget-style thesaurus (*Macquaries* (Bernard, 1986)) as coarse senses or concepts. There are about 1000 categories in a thesaurus, and each category has on average 120 closely related words. A word may be found in more than one category if it has multiple meaning. They use a simple unsupervised algorithm to determine the vector of words that tend to co-occur with each concept and the corresponding strength of association (a measure of how strong the tendency to co-occur is). The target word  $u$  will be assigned one DPC for each of the concepts that

list  $u$ . Below are example DPCs of the two concepts pertaining to *bank*:<sup>4</sup>

DPC(“fin. inst.”):  
*money,1000; boat,32; bond,705; fish,0; ...*  
 DPC(“river bank”):  
*money,5; boat,863; bond,0; fish,948; ...*

The distance between two words  $u, v$  is determined by calculating the closeness of each of the DPCs of  $u$  to each of DPCs of  $v$ , and the closest DPC-pair distance is chosen.

Mohammad and Hirst (2006) show that their approach performs better than other strictly corpus-based approaches that they experimented with. However, all those experiments were on word-pairs that were listed in the thesaurus. Their approach is not applicable otherwise. In Sections 3 and 4 we show how cosine–log-likelihood-ratio (or any comparable distributional measure) can be combined with the Mohammad and Hirst DPCs to form a hybrid approach that is not limited to the vocabulary of a lexical resource.

Erk and Padó (2008) proposed a way of representing a word sense in context by biasing the target word’s DP according to the context surrounding a target (specific) occurrence of the target word. They use dependency relations and selectional preferences of the target word and combine multiple DPs of words appearing in the context of the target occurrence, in a manner so as to give more weight to words co-occurring with both the target word and the target occurrence’s context words. The advantage of this approach is that it does not rely on a thesaurus or WordNet. Its disadvantage is that it relies on dependency relations and selectional preferences information, and that the context information it uses in order to determine the word sense is quite limited (only the words surrounding a single occurrence of the and hence the representation of that sense might not be sufficiently accurate. Their approach effectively assumes that each occurrence of a word has a unique sense.

### 3 Distributional Measures with Soft Semantic Constraints

Traditional distributional profiles of words (DPW) give word–word co-occurrence frequencies. For example,  $DPW(u)$  gives the number of times

<sup>4</sup>The relatively large co-occurrence frequency values for DPCs as compared to DPWs is because a concept can be referred to by many words (on average 100).

the target word  $u$  co-occurs with with all other words:<sup>5</sup>

$DPW(u)$ :  
 $w_1, f(u, w_1); w_2, f(u, w_2); w_3, f(u, w_3); \dots$

where  $f$  stands for co-occurrence frequency (and can be generalized to stand for any strength of association (SoA) measure such as the log-likelihood ratio). Mohammad and Hirst create concept–word co-occurrence vectors, “distributional profiles of concepts” (DPCs), from non-annotated corpus. For example,  $DPC(c)$  gives the number of times the concept (thesaurus category)  $c$  co-occurs with all the words in a corpus.

$DPC(c)$ :  
 $w_1, f(c, w_1); w_2, f(c, w_2); w_3, f(c, w_3); \dots$

A target word  $u$  that appears under thesaurus concepts  $c_1, \dots, c_n$  would be assigned to  $DPC(c_1), \dots, DPC(c_n)$ . Therefore, if a target word  $v$  also appears under some same concept  $c$ , the DPCs of  $u$  and  $v$  would be indistinguishable.

We propose the creation of distributional profiles of word senses ( $DPWS(u_c)$ ) that approximate the SoA of the target word  $u$ , when used in sense  $c$ , with each of the words in the corpus:

$DPWS(u_c)$ :  
 $w_1, f(u_c, w_1); w_2, f(u_c, w_2); w_3, f(u_c, w_3); \dots$

In order to get exact counts, one needs sense-annotated data. However, such data is expensive to create, and is scarce. Therefore, we propose estimating these counts from the DPW and DPC counts:

$$f(u_c, w_i) = p(c|w_i) \times f(u, w_i) \quad (1)$$

where the conditional probability  $p(c|w_i)$  is calculated from the co-occurrence frequencies in DPCs; and the co-occurrence count  $f(u, w_i)$  is calculated from DPWs. If the target word is not in the thesaurus’s vocabulary, then we assume uniform distribution over all concepts, and in practice use a single sense, and take the conditional probability to be 1. Since the method takes sense-proportional co-occurrence counts, we will refer to this method as the **hybrid-sense-proportional-counts method** (or, **hybrid-prop** for short).

<sup>5</sup>The dimensions of the DP co-occurrence vector can be defined arbitrarily, and do not have to correspond to the words in the vocabulary. The most notable alternative representation is the Latent Semantic Analysis and its variants (Landauer et al., 1998; Finkelstein et al., 2002; Budiu et al., 2006).

For example, below is the DPWS of *bank* in the “financial institution” sense, calculated from its DPW and DPCs:

DPW(*bank*):  
*money*,100; *boat*,80; *bond*,70; *fish*,77; ...

DPC(“fin. inst.”):  
*money*,1000; *boat*,32; *bond*,705; *fish*,0; ...

DPC(“river bank”):  
*money*,5; *boat*,863; *bond*,0; *fish*,948; ...

DPWS(*bank* “fin. inst.”):  
*money*,( $\frac{1000}{1000+5} \times 100$ ); *boat*,( $\frac{32}{32+863} \times 80$ );  
*bond*,( $\frac{705}{705+0} \times 70$ ); *fish*,( $\frac{0}{0+948} \times 77$ ); ...

Once the DPWS are calculated, any counts-based SoA and distance measures can be applied. For example, in this work we use log-likelihood ratio (Dunning, 1993) to determine the SoA between a word sense and co-occurring words, and cosine to determine the distance between two DPWS’s log likelihood vectors (McDonald, 2000). We also contrast this measure with cosine of conditional probabilities vectors. Given two target words, we determine the distance between each of their DPWS pairings and the closest DPWS-pair distance is chosen.

### 3.1 The hybrid-sense-filtered-counts method

Since the DPCs are created in an unsupervised manner, they are expected to be somewhat noisy. Therefore, we also experimented with a variant of the method proposed above, that simply makes use of whether the conditional probability  $p(c|w_i)$  is greater than 0 or not:

$$f(u_c, w_i) = \begin{cases} f(u, w_i) & \text{If } p(c|w_i) > 0 \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

Since this method essentially filters out collocates that are likely not relevant to the target sense  $c$  of the target word  $u$ , we will refer to this method as the **hybrid-sense-filtered-counts method** (or, just **hybrid-filt** for short). Below is an example hybrid-filtered DPWS of *bank* in the “financial institution” sense:

DPWS(*bank* “fin. inst.”):  
*money*,100); *boat*,80; *bond*,70; ...

Note that the collocate *fish* is now filtered out, whereas *bank*’s co-occurrence counts with *money*, *boat*, and *bond* are left as is (and not sense-proportionally attenuated).

## 4 Evaluation

We evaluated various methods on the task of ranking word pairs in order of semantic distance. These methods included our sense-biased methods as well as several baselines: the Mohammad and Hirst (2006) DPC-based methods, the traditional word-based distributional similarity methods, and several Latent Semantic Analysis (LSA)-based methods. We used three testsets and their corresponding human judgment gold standards: (1) the Rubenstein and Goodenough (1965) set of 65 noun pairs—denoted **RG-65**; (2) the WordSimilarity-353 (Finkelstein et al., 2002) set of 353 noun pairs (which include the RG-65 pairs) of which we discarded of one repeating pair—denoted **WS-353**; and (3) the Resnik and Diab (2000) set of 27 verb pairs—denoted **RD-00**.

### 4.1 Corpora and Pre-processing

We generated distributional profiles (DPWs and DPCs) from the *British National Corpus (BNC)* (Burnard, 2000), which is a balanced corpus. We lowercased the characters, and stripped numbers, punctuation marks, and any SGML-like syntactic tags, but kept sentence boundary markers. The BNC contained 102,100,114 tokens of 546,299 types (vocabulary size) after tokenization. For the verb set, we also lemmatized this corpus.

We considered two words as co-occurring if they occurred in a window of  $\pm 5$  words from each other. We stoplisted words that co-occurred with more than 2000 word types.

### 4.2 Results

The Spearman rank correlations of the automatic rankings of the RG-65, WS353, and RD-00 testsets with the corresponding gold-standard human rankings is listed in Table 1.<sup>6</sup> The higher the Spearman rank correlation, the more accurate is the distance measure.

#### 4.2.1 Results on the RG-65 testset

**Baselines.** We replicated the traditional word-based distributional distance measure using cosine of vectors (DPs) containing conditional probabilities (**word-cos-cp**). Its rank correlation of .53 is close to the correlation of .54 reported in Mohammad and Hirst (2006), hereafter MH06. We replicated the MH06 concept-based approach

<sup>6</sup>Certain experiments were not pursued as they were redundant in supporting our claims.

Method	RG-65	WS-353	RD-00
<b>Baselines (replicated):</b>			
<i>Traditional distributional measures</i>			
word-cos-cp	.53	.31	.46
word-cos-ll	.70	<b>.54</b>	.51
word-cos-pmi	.62	.43	.57
<i>Mohammad and Hirst methods and variants</i>			
concept-cos-cp	.62	.38	.41
concept*-cos-cp	.65	.33	.43
concept-cos-ll	.60	.37	.43
concept*-cos-ll	.64	.25	.27
concept*-cos-pmi	.40	.19	.28
<i>Other (LSA and variants)</i>			
LSA	.56	.47	.55
GLSA-cos-pmi	.18	n.p.	n.p.
GLSA-cos-ll	.47	n.p.	.29
<b>New methods:</b>			
hybrid-prop-cos-ll	.72	.49	.53
hybrid-prop*-cos-ll	.69	.46	.45
hybrid-filt-cos-ll	.73	<b>.54</b>	.38
hybrid-filt*-cos-ll	<b>.77</b>	<b>.54</b>	.39
hybrid-prop*-cos-pmi	.58	.43	<b>.71</b>
hybrid-filt*-cos-pmi	.61	.42	.64

Table 1: Spearman rank correlation on RG-65, WS-353, and RD-00 testsets, trained on BNC. ‘\*’ indicates the use of a smaller bootstrapped concept–word co-occurrence matrix. ‘n.p.’ indicates that the experiment was not pursued.

(**concept-cos-cp**), and its bootstrapped variant that uses a smaller concept–word co-occurrence matrix (**concept\*-cos-cp**). The latter yielded a correlation score .65, close to the .69 reported in MH06.

We also experimented with cosine of PMI vectors (**word-cos-pmi**) which obtained a correlation of .62. Log likelihood ratios (**word-cos-ll**) gave best results among the baseline methods (.70), and so we it more often in the implementations of our hybrid method.

We conducted experiments with LSA and its GLSA variants (Budiu et al., 2006) as additional baselines. A limited vocabulary of the 33,000 most frequent words in the BNC and all test words was used in these experiments. (A larger vocabulary was computationally expensive and 33,000 is also the vocabulary size used by Budiu et al. (2006) in their LSA experiments.)

**New Methods:** The hybrid method variants proposed in this paper (**hybrid-prop-cos-ll** and **hybrid-filt-cos-ll**) were the best performers on the RG-65 test set. Particularly, they performed better than both the traditional word-distance measures (**word-cos-ll**), and our concept-based methods—variants of the MH06 method that are used with likelihood ratios (**concept-cos-ll**, **concept\*-cos-**

**ll**). The -pmi methods were all poorer performers than their -ll counterparts. The -pmi hybrid variants obtained higher scores than the concept-based ones, but almost the same scores as the word-based ones.

#### 4.2.2 Results on WS-353 and RD-00 testsets

On WS-353, all our hybrid methods outperformed their concept counterparts, and were on par with their word-based counterparts. On RD-00, **word-cos-pmi** outperformed all other word-based methods, and the hybrid -pmi methods were best performers with scores of .64 and .71. Our word-cos-ll, hybrid-prop-cos-ll, and the two hybrid pmi results on RD-00 are better than any non-WordNet results reported by Resnik and Diab (2000), including their syntax-informed methods—the variants of Lin (“distrib”, .43) and Dorr (“LCS”, .39). In fact, our hybrid\*-prop-cos-pmi and hybrid\*-filt-cos-pmi results reach correlation levels of the WordNet-based methods reported there (.66–.68). Also, on WS-353, our hybrid sense-filtered variants and word-cos-ll obtained a correlation score higher than published results using WordNet-based measures (Jarmasz and Szpakowicz, 2003) (.33 to .35) and Wikipedia-based methods (Ponzetto and Strube, 2006) (.19 to .48); and very close to the results obtained by thesaurus-based (Jarmasz and Szpakowicz, 2003) (.55) and LSA-based methods (Finkelstein et al., 2002) (.56).

The lower correlation scores of all measures on the WS-353 test set are possibly due to it having politically biased word pairs (examples include: *Arafat–peace*, *Arafat–terror*, *Jerusalem–Palestinian*) for which BNC texts are likely to induce low correlation with the human raters of WS-353. This testset also has disproportionately many terms from the news domain.

The concept methods performed poorly on WS-353 partly because many of the target words do not exist in the thesaurus. For instance, there were 17 such word types that occurred in 20 WS-353 testset word pairs. When excluding these pairs, concept-cos-cp goes up from .38 to .45, and concept\*-cos-pmi from .19 to .24. Interestingly, results of the hybrid methods show that they were largely unaffected by the out-of-vocabulary problem on the WS-353 dataset.

On the verbs dataset RD-00, while hybrid-prop-cos-ll fared slightly better than word-cos-ll, using the smaller matrix seemed to hurt performance of

hybrid\*-prop-cos-ll compared to word-cos-ll. But results suggest that the -pmi methods might serve as a better measure than -ll for verbs, although this claim should be tested more rigorously.

Human judgments of semantic distance are less consistent on verb-pairs than on noun-pairs, as reflected in inter-rater agreement measures in Resnik and Diab (2000) and others). Thus, not surprisingly, the scores of almost all measures are lower for the verb data than the RG-65 noun data.

## 5 Discussion

The hybrid methods proposed in this paper obtained higher accuracies than all other methods on the RG-65 testset (all of whose words were in the published thesaurus), and on the RD-00 testset, and their performance was at least respectable on the WS-353 testset (many of whose words were not in the published thesaurus). This is in contrast to the concept-distance methods which suffer greatly when the target words are not in the lexical resource (here, the thesaurus) they rely on, even though these methods can make use of co-occurrence information of words not in the thesaurus with concepts from the thesaurus.

Amongst the two hybrid methods proposed, the **sense-filtered-counts** method performed better using the smaller bootstrapped concept-word co-occurrence matrix whereas the sense-proportional method performed better using the larger concept-word co-occurrence matrix. We believe this is because the bootstrapping method proposed in Mohammad and Hirst (2006) has the effect of resetting to 0 the small co-occurrence counts. The noise from these small co-occurrence counts affects the sense-filtered-counts method more adversely (since any non-zero value will cause the inclusion of the corresponding collocate's full co-occurrence count) and so the bootstrapped matrix is more suitable for this method.

The results also show that the cosine of log-likelihood ratios method mostly performs better than cosine of conditional probabilities and the pmi methods on the noun sets. This further supports the claim by Dunning (1993) that log-likelihood ratio is much less sensitive than pmi to low counts. Interestingly, on the verb set, the pmi methods, and especially hybrid\*-prop-cos-pmi, did extremely well. Further investigation is needed in order to determine if pmi is indeed more suitable for verb semantic similarity, and why.

## 6 Conclusion

Traditional distributional similarity conflates co-occurrence information pertaining to the many senses of the target words. Mohammad and Hirst (2006) show how distributional measures can be used to compute distance between very coarse word senses or concepts (thesaurus categories), and even obtain better results than traditional distributional similarity. However, their method requires that the target words be listed in the thesaurus, which is often not the case for domain-specific terms and named entities. In this paper, we proposed hybrid methods (**hybrid-sense-filtered-counts** and **hybrid-sense-proportional-counts**) that combine word-word co-occurrence information (traditional distributional similarity) with word-concept co-occurrence information (Mohammad and Hirst, 2006), with soft constraints in such a manner that the method makes use of information encoded in the thesaurus when available, and degrades gracefully if the target word is not listed in the thesaurus. Our method generates word-sense-biased distributional profiles (DPs) from non-annotated corpus-based word-based DPs and coarser-grained aggregated thesaurus-based "concept DPs" (DPCs). We showed that the hybrid method correlates with human judgments of semantic distance in most cases better than any of the other methods we replicated.

We are now interested in improving semantic distance measures for verb-verb, adjective-adjective, and cross-part-of-speech pairs, by exploiting specific information pertaining to these parts of speech in lexical resources in addition to purely co-occurrence information.

## Acknowledgments

We thank Mona Diab for her help with her verb test set, Raluca Budiu for her help and clarifications regarding the GLSA method and its implementation details, and the anonymous reviewers for their valuable feedback. This work was supported, in part, by the National Science Foundation under Grant No. IIS-0705832, and in part, by the Human Language Technology Center of Excellence. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsor.

## References

- Eneko Agirre and Oier Lopez de Lacalle Lekuona. 2003. Clustering WordNet word senses. In *Proceedings of the 1st International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, Borovets, Bulgaria.
- Satanjeev Banerjee and Ted Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*, pages 805–810, Acapulco, Mexico.
- John R. L. Bernard, editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.
- Raluca Budiú, Christiaan Royer, and Peter Pirolli. 2006. Modeling information scent: A comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. In *Proceedings of RIAO'07*, Pittsburgh, PA.
- Lou Burnard. 2000. *Reference Guide for the British National Corpus*. Oxford University Computing Services, Oxford, England, world edition edition.
- James R. Curran. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.
- Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of cooccurrence probabilities. *Machine Learning*, 34(1–3):43–69.
- Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 897–906, Honolulu, HI.
- Lev Finkelstein, Evgeniy Gábrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppín. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.
- John R. Firth. 1957. A synopsis of linguistic theory 193055. *Studies in Linguistic Analysis*, (special volume of the Philological Society):132. Distributional Hypothesis.
- Zellig S. Harris. 1940. Review of Louis H. Gray, foundations of language (New York: Macmillan, 1939). *Language*, 16(3):216–231.
- Graeme Hirst and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *NLE*, 11(1):87–111.
- Mario Jarmasz and Stan Szpakowicz. 2003. Roger's Thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-2003)*, pages 212–219, Borovets, Bulgaria.
- Thomas Landauer, Peter Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259 – 284.
- Dekang Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, page 296304, San Francisco, CA.
- Adam Lopez. 2008. Statistical machine translation. *ACM Computing Surveys*, 40(3):149.
- Diana McCarthy. 2006. Relating WordNet senses for word sense disambiguation. In *Proceedings of the European Chapter of the Association for Computational Linguistics Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 17–24, Trento, Italy.
- S. McDonald. 2000. *Environmental determinants of lexical processing effort*. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.
- Saif Mohammad and Graeme Hirst. 2006. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of EMNLP*.
- Saif Mohammad, Iryna Gurevych, Graeme Hirst, and Torsten Zesch. 2007. Cross-lingual distributional profiles of concepts for measuring semantic distance. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP/CoNLL-2007)*, pages 571–580, Prague, Czech Republic.
- Saif Mohammad. 2008. *Measuring Semantic Distance using Distributional Profiles of Concepts*. Ph.D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.
- Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 46–51, Boston, Massachusetts.
- Roberto Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association*, pages 105–112, Sydney, Australia.
- Siddharth Patwardhan and Ted Pedersen. 2006. Using WordNet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of Making Sense of Sense EACL Workshop*, pages 1–8.
- Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2006)*, pages 192–199, New York, NY.
- Philip Resnik and Mona Diab. 2000. Measuring verb similarity. In *22nd Annual Meeting of the Cognitive Science Society (COGSCI2000)*, Philadelphia, PA.
- Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *JAIR*, 11:95–130.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 1015–1021, Geneva, Switzerland.