

# Measuring Semantic Distance using Distributional Profiles of Concepts

Saif Mohammad\*  
Institute of Advanced Computer  
Studies, University of Maryland

Graeme Hirst\*\*  
Department of Computer Science,  
University of Toronto

*Automatic measures of semantic distance can be classified into two kinds: (1) those, such as WordNet, that rely on the structure of manually created lexical resources and (2) those that rely only on co-occurrence statistics from large corpora. Each kind has inherent strengths and limitations. Here we present a hybrid approach that combines corpus statistics with the structure of a Roget-like thesaurus to gain the strengths of each while avoiding many of their limitations. We create distributional profiles (co-occurrence vectors) of coarse thesaurus concepts, rather than words. This allows us to estimate the distributional similarity between concepts, rather than words. We show that this approach can be ported to a cross-lingual framework, so as to estimate semantic distance in a resource-poor language by combining its text with a thesaurus in a resource-rich language. Extensive experiments, both monolingually and cross-lingually, on ranking word pairs in order of semantic distance, correcting real-word spelling errors, and solving word-choice problems show that these distributional measures of concept distance markedly outperform traditional distributional word-distance measures and are competitive with the best WordNet-based measures.*

## 1. Introduction

**Semantic distance** is a measure of how close or distant two units of language are, in terms of their meaning. The units of language may be words, phrases, sentences, paragraphs, or documents. For example, the nouns *dance* and *choreography* are closer in meaning than the nouns *clown* and *bridge*. These units of language, especially words, may have more than one possible meaning. However, their context may be used to determine the intended senses. For example, *star* can mean both CELESTIAL BODY and CELEBRITY; however, *star* in the sentence below refers only to CELESTIAL BODY and is much closer to *sun* than to *famous*:

(1) *Stars are powered by nuclear fusion.*

Thus, semantic distance between words in context is in fact the distance between word senses or concepts. (We use the terms *word senses* and *concepts* interchangeably here, although later on we will make a distinction. Throughout this paper, example words will be written in italics, as in the example sentence above, whereas example senses or concepts will be written in small capitals.)

---

\* E-mail: saif@umiacs.umd.edu  
\*\* Email: gh@cs.toronto.edu

Two classes of automatic semantic distance measures exist. **Lexical-resource-based measures of concept-distance**, such as those of Jiang and Conrath (1997), Leacock and Chodorow (1998), and Resnik (1995), rely on the structure of a knowledge source, such as WordNet, to determine the distance between two concepts defined in it. **Distributional measures of word-distance**, such as cosine and  $\alpha$ -skew divergence (Lee 2001), rely on the **distributional hypothesis**, which states that two words are semantically close if they tend to occur in similar contexts (Firth 1957; Harris 1968). These measures rely simply on text and can give the distance between any two words that occur at least a few times.

However, both these approaches have significant limitations (described in detail in Section 3). In this paper (Section 4), we present a new hybrid approach that combines the co-occurrence statistics of a distributional approach with the information in a lexical resource. We will refer to this new class of measures as **distributional measures of concept-distance**. We also show how this approach can be ported to a cross-lingual framework (Section 6), which has two additional benefits: (1) Semantic distance problems in a resource-poor language can be solved by combining its texts with a lexical resource from a resource-rich language; (2) Cross-lingual semantic distance is useful when working on natural language problems that inherently involve two or more languages, such as machine translation and multilingual information retrieval. We perform extensive experiments, both monolingually (Section 5) and cross-lingually (Section 7), and show that this new method is markedly better than others.

## 2. Background

### 2.1 When are two terms considered semantically close?

Humans consider two concepts to be semantically close if there is a sharing of some meaning. More formally, two concepts are semantically close if there is a **lexical semantic relation** between the concepts. According to Cruse (1986), a lexical semantic relation is the relation between **lexical units**—a surface form along with a sense. As Cruse points out, the number of semantic relations that bind concepts is innumerable but certain relations, such as hyponymy, meronymy, antonymy, and troponymy, are more systematic and have enjoyed more attention in the linguistics community. However, as Morris and Hirst (2004) point out, these relations are far out-numbered by others which they call **non-classical relations**. A few of the kinds of non-classical relations they observed included positive qualities (BRILLIANT, KIND), commonly co-occurring words (locations such as HOMELESS, SHELTER; problem–solution pairs such as DRUGS, REHABILITATION).

### 2.2 Semantic relatedness and semantic similarity

Semantic distance is of two kinds: **semantic similarity** and **semantic relatedness**. The former is a subset of the latter, but the two may be used interchangeably in certain contexts, making it even more important to be aware of their distinction. Two concepts are considered to be semantically similar if there is a hyponymy (hypernymy), antonymy, or troponymy relation between them. Two concepts are considered to be semantically related if there is any lexical semantic relation between them—classical or non-classical.

Semantically similar concepts tend to share a number of common properties. For example, consider APPLES and BANANAS. They are both hyponyms of FRUIT. They are both edible, they grow on trees, they have seeds, etc. Another example of a semantically

**Table 1**

Word-pair datasets that have been manually annotated with distance values. Pearson’s correlation was used to determine inter-annotator correlation (last column). Those used for experiments reported in this paper are marked in bold. “n.r.” stands for “not reported”.

<b>Dataset</b>	<b>Year</b>	<b>Language</b>	<b># pairs</b>	<b>PoS</b>	<b># subjects</b>	<b>Correlation</b>
<b>Rubenstein and Goodenough</b>	1965	English	65	N	51	n.r.
Miller and Charles	1991	English	30	N	n.r.	.90
Resnik and Diab	2000	English	27	V	n.r.	.76 and .79
Finkelstein	2002	English	153	N	13	n.r.
Finkelstein	2002	English	200	N	16	n.r.
<b>Gurevych</b>	2005	German	65	N	24	.81
<b>Zesch and Gurevych</b>	2006	German	350	N, V, A	8	.69

similar pair is DOCTOR and SURGEON. The concept of a DOCTOR is a hypernym of SURGEON. Therefore, they share the properties associated with a DOCTOR.

On the other hand, semantically related concepts may not have many properties in common, but have at least one classical or non-classical lexical relation between them which lends them the property of being semantically close. For example, DOOR and KNOB are semantically related as one is the meronym of the other (i.e., stands in the part-of relation). The concept pair, DOCTOR and SURGEON is in addition to semantically related (as well as being semantically similar) as one is the hyponym of the other. Example pairs considered semantically related due to non-classical relations include SURGEON–SCALPEL and TREE–SHADE. Note that semantic similarity entails semantic relatedness but the converse need not be true.

### 2.3 Can humans estimate semantic distance?

Many will agree that humans are adept at estimating semantic distance, but consider the following questions. How strongly will two people agree or disagree on distance estimates? Will the agreement vary over different sets of concepts? In our minds, is there a clear distinction between related and unrelated concepts or are concept-pairs spread across the whole range from synonymous to unrelated?

Some of the earliest work that begins to answer these questions is by Rubenstein and Goodenough (1965a). They conducted quantitative experiments with human subjects (51 in all) who were asked to rate 65 English word pairs on a scale from 0.0 to 4.0 as per their semantic distance. The word pairs chosen ranged from almost synonymous to unrelated. However, they were all noun pairs and those that were semantically close were semantically similar; the dataset did not contain word pairs that were semantically related but not semantically similar. The subjects repeated the annotation after two weeks and the new distance values had a Pearson’s correlation  $r$  of 0.85 with the old ones. Miller and Charles (1991) also conducted a similar study on 30 word pairs taken from the Rubenstein-Goodenough pairs. These annotations had a high correlation ( $r = 0.97$ ) with the mean annotations of Rubenstein and Goodenough (1965a). Resnik (1999) repeated these experiments and found the inter-annotator correlation ( $r$ ) to be 0.90. Finkelstein (2002) asked human judges to rank two sets of noun pairs (153 pairs and 200 pairs) in order of semantic distance. However, this dataset has certain politically biased word pairs, such as *Arafat–peace*, *Arafat–terror*, *Jerusalem–Israel*, *Jerusalem–Palestinian*, and so there might be less human agreement on ranking this data.

Resnik and Diab (2000) conducted annotations of 48 verb pairs and found inter-annotator correlation ( $r$ ) to be 0.76 when the verbs were presented without context and 0.79 when presented in context. Gurevych (2005) and Zesch et al. (2007) asked native German speakers to mark two different sets of German word pairs with distance values. Set 1 was a German translation of the Rubenstein and Goodenough (1965a) dataset. It had 65 noun–noun word pairs. Set 2 was a larger dataset containing 350 word pairs made up of nouns, verbs, and adjectives. The semantically close word pairs in the 65-word set were mostly synonyms or hypernyms (hyponyms) of each other, whereas those in the 350-word set had both classical and non-classical relations with each other. Details of these **semantic distance benchmarks** are summarized in Table 1. Inter-subject correlations (last column in Table 1) are indicative of the degree of ease in annotating the datasets.

The high correlation values suggest that humans are quite good and consistent at estimating semantic distance of noun-pairs; however, annotating verbs and adjectives and combinations of parts of speech is harder. This also means that estimating semantic relatedness is harder than estimating semantic similarity. It should be noted here that even though the annotators were presented with word-pairs and not concept-pairs, it is reasonable to assume that they were annotated as per their closest senses. For example, given the noun pair *bank* and *interest*, most if not all will identify it as semantically related even though both words have more than one sense and many of the sense–sense combinations are unrelated (for example, the RIVER BANK sense of *bank* and the SPECIAL ATTENTION sense of *interest*).

Apart from proving that humans can indeed estimate semantic distance, these datasets act as “gold standards” to evaluate automatic distance measures. However, lack of large amounts of data from human subject experimentation limits the reliability of this mode of evaluation. Therefore automatic distance measures are also evaluated by their usefulness in natural language tasks such as correcting real-word spelling errors (Budanitsky and Hirst 2006) and solving word-choice problems (Turney 2001). We evaluate our distributional concept-distance measures both intrinsically through ranking human-judged word pairs in order of semantic distance as well as extrinsically through natural language tasks such as correcting spelling errors and solving word-choice problems.

#### 2.4 The anatomy of a distributional measure of semantic distance

Even though there are numerous distributional measures, many of which may seem dramatically different from each other, all of them do the following: (1) choose a unit of co-occurrence (e.g., word, word–syntactic-relation combination); (2) choose a measure of strength of association (SoA) of the co-occurrence unit with the target word (e.g., conditional probability, pointwise mutual information); (3) represent the target words by vectors or points in the co-occurrence space (and possibly apply dimension reduction);<sup>1</sup> and (4) calculate the distance between the target vectors using a suitable distributional measure (e.g., cosine, Euclidean distance). While any of the measures of vector distance may be used with any of the measures of strength of association, in practice only certain combinations are used (see Table 2) and certain other combinations

---

<sup>1</sup> The co-occurrence space is a hyper-dimensional space where each dimension is a unique co-occurrence unit. If words are used as co-occurrence units, then this space has  $|V|$  dimensions, where  $V$  is the vocabulary.

**Table 2**

Measures of vector distance, measures of strength of association, and standard combinations. Those used for experiments reported in this paper are marked in bold.

Measures of DP distance	Measures of strength of association (SoA)
$\alpha$ -skew divergence ( <b>ASD</b> )	$\phi$ coefficient (Phi)
<b>cosine (Cos)</b>	<b>conditional probability (CP)</b>
Dice coefficient (Dice)	cosine (Cos)
Euclidean distance ( $L_2$ norm)	Dice coefficient (Dice)
Hindle’s measure (Hin)	log likelihood ration (LLR)
Kullback-Leibler divergence (KLD)	odds ratio (Odds)
Manhattan distance ( $L_1$ norm)	<b>pointwise mutual information (PMI)</b>
<b>Jensen–Shannon divergence (JSD)</b>	Yule’s coefficient (Yule)
<b>Lin’s measure (Lin)</b>	

---

**Standard combinations**


---

$\alpha$ -skew divergence— $\phi$ coefficient ( <b>ASD–CP</b> )
<b>cosine—conditional probability (Cos–CP)</b>
Dice coefficient—conditional probability (Dice–CP)
Euclidean distance—conditional probability ( $L_2$ norm–CP)
Hindle’s measure—pointwise mutual information (Hin–PMI)
Kullback-Leibler divergence—conditional probability (KLD–CP)
Manhattan distance—conditional probability ( $L_1$ norm–CP)
<b>Jensen–Shannon divergence—conditional probability (JSD–CP)</b>
<b>Lin’s measure—pointwise mutual information (Lin–PMI)</b>

---

may not be meaningful, for example, Kullback-Leibler divergence with  $\phi$  coefficient. We will refer to these co-occurrence vectors as the **distributional profile (DP)** of the target words. Below is a contrived, but plausible, example DP of the target word *fusion*:

FUSION: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, ...

It shows that *fusion* has a strong tendency to co-occur with words such as *heat*, *hydrogen*, and *energy*. The values are the pointwise mutual information between the target and co-occurring words.

All experiments in this paper use simple word co-occurrences, and standard combinations of vector distance and measure of association. To avoid clutter, instead of referring to a distributional measure by its measure of vector distance and measure of association (for example,  $\alpha$ -skew divergence—conditional probability), we will refer to it simply by the measure of vector distance (in this case,  $\alpha$ -skew divergence). The measures used in our experiments are  $\alpha$ -skew divergence (**ASD**) (Lee 2001), cosine (**Cos**) (Schütze and Pedersen 1997), Jensen-Shannon divergence (**JSD**) (Manning and Schütze 2008), and that proposed by Lin (1998a) (**Lin**). Jensen–Shannon divergence and  $\alpha$ -skew divergence calculate the difference in distributions of words that co-occur with the targets. Lin’s distributional measure follows from his information-theoretic definition of similarity (Lin 1998b).

### 3. Limitations of semantic distance measures

Lexical-resource-based concept-distance measures and distributional word-distance measures each have certain uniquely attractive features: resource-based measures can capitalize on manually encoded lexical semantic relations, whereas distributional approaches are widely applicable because they need only raw text (and maybe some shallow syntactic processing). However, they both also have certain limitations.

#### 3.1 Limitations of lexical-resource-based concept-distance measures

Resource-based measures are only as good as the lexical resource on which they rely.

**3.1.1 Lack of high-quality WordNet-like knowledge sources.** Ontologies, wordnets, and semantic networks are available for a few languages such as English, German, and Hindi. Creating them requires human experts and it is time intensive. Thus, for most languages, we cannot use resource-based measures simply due to the lack of high-quality large-coverage wordnet-like resources. Further, updating a resource is again expensive and there is usually a lag between the current state of language usage/comprehension and the lexical resource representing it.

On the other hand, distributional measures require only text. Large corpora, billions of words in size, may now be collected by a simple web crawler. Large corpora of more-formal writing are also available (for example, the *Wall Street Journal* or the *American Printing House for the Blind (APHB)* corpus). This makes distributional measures very attractive.

**3.1.2 Poor estimation of semantic relatedness.** The most widely used WordNet-based measures rely only on its extensive  $\hat{is}$ -a hierarchy. This is because networks of other lexical-relations such as meronymy are much less developed. Further, the networks for different parts of speech are not well connected. Thus, even though resource-based measures are successful at estimating semantic similarity between nouns, they are poor at estimating semantic relatedness—especially in pairs other than noun–noun. Also, as Morris and Hirst (2004) pointed out, a large number of terms have a non-classical relation between them and are semantically related (not semantically similar). On the other hand, distributional measures can be used to determine both semantic relatedness and semantic similarity (Mohammad and Hirst 2007).

**3.1.3 Inability to cater to specific domains.** Given a concept pair, measures that rely only on a network and no text, such as Rada et al. (1989), give just one distance value. However, two concepts may be very close in a certain domain but not so much in another. For example, SPACE and TIME are close in the domain of quantum mechanics but not so much in most others. Resources created for specific domains do exist; however, they are rare. Some of the more successful WordNet-based measures, such as that of Jiang and Conrath (1997), rely on text as well, and do indeed capture domain-specificity to some extent, but the distance values are still largely affected by the underlying network, which is not domain-specific. On the other hand, distributional measures rely primarily (if not completely) on text and large amounts of corpora specific to particular domains can easily be collected.

### 3.2 Limitations of corpus-based word-distance measures

**3.2.1 Conflation of word senses.** The distributional hypothesis (Firth 1957) states that words that occur in similar contexts tend to be semantically close. But most frequently used words have more than one sense; and a word in each of its senses is likely to co-occur with different sets of words. For example, *bank* in the FINANCIAL INSTITUTION sense is likely to co-occur with *interest*, *money*, *accounts*, and so on, whereas when used in the RIVER BANK sense will co-occur with such as *river*, *erosion*, and *silt*. Thus, a distributional measure will give a score that is some form of a dominance-based average of the distances between their senses.

However, in most natural language applications, including spelling correction, information retrieval, and text summarization, one requires the semantic distance between the intended senses of the target words. Since words that occur together in text tend to refer to senses that are closest in meaning to one another, this often tends to be the distance between the closest senses of the two target words. Thus, distributional word-distance measures are expected to perform poorly in the face of word sense ambiguity. WordNet-based measures do not suffer from this problem as they give distance between concepts, not words.

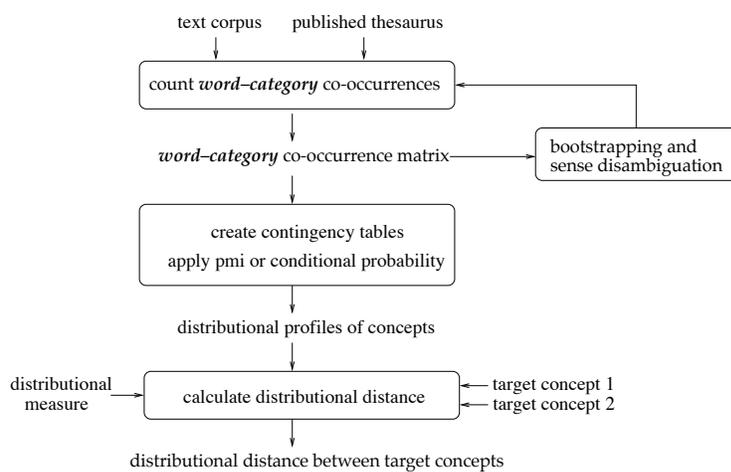
**3.2.2 Data sparseness.** Since Zipf's law seems to hold even for the largest of corpora, there will always be words that occur too few times for distributional measures to accurately estimate their distance with other words. On the other hand, a large number of relatively obscure words may be listed in high-coverage resources such as WordNet (WordNet has more than 155,000 unique tokens). Of course, manually created resources are also lacking in a number of word-types. However, they tend to have groupings of words into coarse concepts. This allows even corpus-based approaches to determine properties of these coarse concepts through occurrences of the more frequent members of a concept. In Section 4, we will propose a hybrid method of semantic distance that does exactly that using the categories in a published thesaurus.

### 3.3 Space requirements

As applications for linguistic distance become more sophisticated and demanding, it becomes attractive to pre-compute and store the distance values between all possible pairs of words or senses. However both WordNet-based and distributional measures have large space requirements to do this, requiring matrices of size  $N \times N$ , where  $N$  is very large. In case of distributional measures,  $N$  is the size of the vocabulary (at least 100,000 for most languages). In case of WordNet-based measures,  $N$  is the number of senses (81,000 just for nouns). Given that these matrices tend to be sparse<sup>2</sup> and that computational capabilities are continuing to improve, this limitation may not seem hugely problematic, but as we see more and more natural language applications in embedded systems and hand-held devices, such as cell phones, iPods, and medical equipment, available memory becomes a serious constraints.

---

<sup>2</sup> Even though WordNet-based and distributional measures give non-zero similarity and relatedness values to a large number of term pairs (concept pairs and word pairs), values below a suitable threshold can be reset to 0.



**Figure 1**  
An overview of the distributional concept-distance approach.

#### 4. Distributional Measures of Concept-Distance

We now propose a hybrid approach that combines corpus statistics with a published thesaurus (Mohammad and Hirst 2006b; Mohammad et al. 2007). It overcomes, with varying degrees of success, many of the limitations described in Section 3 earlier. Our goal is to gain the performance of resource-based methods and the breadth of distributional methods. The central ideas are these:

- In the lexicographical component of the method, concepts are defined by the category structure of a Roget-style thesaurus.
- In order to avoid data sparseness, the concepts are very coarse-grained.
- The distributional component of the method is based on concepts, not surface strings. We create distributional profiles (co-occurrence vectors) of *concepts*.

The sub-sections below describe our approach in detail. Figure 1 depicts the key steps.

##### 4.1 Published thesaurus

A Roget-style thesaurus classifies all word types into approximately 1000 categories. Words within a category tend to be semantically related to each other. Words with more than one sense are listed in more than one category. Each category has a head word that best represents the meaning of all the words in the category. Some example categories are CLOTHING, HONESTY, and DESIRE. Each category is divided into paragraphs that classify lexical units more finely; however, we do not make use of this information.

We take these thesaurus categories as the coarse-grained concepts of our method. That is, for our semantic distance measure, there are only around 1000 concepts (word-senses) in the world; each lexical unit is a pairing of the surface string with the thesaurus

category in which it appears. This is in stark contrast to using WordNet synsets as senses, which sometimes has been criticized to be much too fine-grained.<sup>3</sup>

Representing the complete vocabulary with only about 1000 concepts helps counter data sparseness issues (limitation described in Section 3.2.2) at the cost of losing the ability to make fine-grained distinctions. This also means that pre-computing a complete concept–concept distance matrix now involves the creation of a matrix approximately only  $1000 \times 1000$  in size—much smaller and roughly .01% the size of matrices required by existing measures—thereby mitigating storage limitations in memory-scarce devices (limitation of Section 3.3).

In our experiments, we use the categories from the *Macquarie Thesaurus* (Bernard 1986). It has 812 categories with around 176,000 word-tokens and 98,000 word-types.

#### 4.2 The distributional hypothesis for concepts

If we apply the distributional hypothesis (Firth 1957; Harris 1968) to word senses (instead of words), then the hypothesis states that words when used in different *senses* tend to keep different “company” (co-occurring words). Therefore, we propose the creation of distributional profiles (DPs) of word senses or concepts, rather than those of words. The closer the distributional profiles of two concepts, the smaller is their semantic distance. Below are example distributional profiles of two senses of STAR:

CELESTIAL BODY: *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...  
 CELEBRITY: *famous* 0.24, *movie* 0.14, *rich* 0.14, *fan* 0.10, ...

It should be noted that creating such distributional profiles of concepts is much more challenging than creating distributional profiles of words, which involve simple word–word co-occurrence counts. (In the next sub-section, we show how these profiles may be estimated without the use of any sense-annotated data). However, once created, any of the many measures of vector distance can be used to estimate the distance between the DPs of two target concepts (just as in the case of traditional word-distance measures, measures of vector distance are used to estimate the distance between the DPs of two target words). For example, here is how cosine is traditionally used to estimate distributional distance between two words.

$$\text{Cos}_{cp}(w_1, w_2) = \frac{\sum_{w \in C(w_1) \cup C(w_2)} (P(w|w_1) \times P(w|w_2))}{\sqrt{\sum_{w \in C(w_1)} P(w|w_1)^2} \times \sqrt{\sum_{w \in C(w_2)} P(w|w_2)^2}} \quad (1)$$

$C(t)$  is the set of words that co-occur (within a certain window) with the word  $t$  in a corpus. The conditional probabilities in the formula are taken from the distributional profiles of words. We adapt the formula to estimate distributional distance between two concepts as shown below:

$$\text{Cos}_{cp}(c_1, c_2) = \frac{\sum_{w \in C(c_1) \cup C(c_2)} (P(w|c_1) \times P(w|c_2))}{\sqrt{\sum_{w \in C(c_1)} P(w|c_1)^2} \times \sqrt{\sum_{w \in C(c_2)} P(w|c_2)^2}} \quad (2)$$

<sup>3</sup> WordNet has more than 117,000 synsets. To counter this fine-grainedness, methods to group synsets into coarser senses have been proposed (Agirre and Lopez de Lacalle Lekuona 2003; Navigli 2006).

$C(x)$  is now the set of words that co-occur with *concept*  $x$  within a pre-determined window. The conditional probabilities in the formula are taken from the distributional profiles of concepts.

If the distance between two words is required, and their intended senses are not known, then the distance between all relevant sense pairs is determined and the minimum is chosen. (This is the heuristic described earlier in Section 3.2.1, and is exactly how WordNet-based measures of concept-distance are used too (Budanitsky and Hirst 2006).) For example, if *star* has the two senses mentioned above and *fusion* has one (let's call it FUSION), then the distance between them is determined by first applying cosine (or any vector distance measure) to the DPs of CELESTIAL BODY and FUSION:

CELESTIAL BODY: *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...  
 FUSION: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, ...

and then applying cosine to the DPs of CELEBRITY and FUSION:

CELEBRITY: *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...  
 FUSION: *heat* 0.16, *hydrogen* 0.16, *energy* 0.13, *bomb* 0.09, *light* 0.09, *space* 0.04, ...

Finally the scores which imply the greatest closeness (least distance) is chosen:

$$\text{distance}(\textit{star}, \textit{fusion}) = \max(\text{Cos}(\text{CELEBRITY}, \text{FUSION}), \text{Cos}(\text{CELESTIAL BODY}, \text{FUSION})) \quad (3)$$

Note that the maximum value is chosen above because cosine is a closeness measure (greater values imply smaller distances). In the case of distance measures, such as  $\alpha$ -skew divergence, the lower of the two values will be chosen.

### 4.3 Generating distributional profiles of concepts

Determining distributional profiles of *concepts* requires information about which words co-occur with which concepts. A direct approach for this requires the text, from which counts are made, to be sense annotated. Since existing labeled data is minimal and manual annotation is far too expensive, indirect means must be used. Below, we present a way to estimate distributional profiles of concepts from raw text, using a published thesaurus (the concept inventory) and a bootstrapping algorithm.

**4.3.1 Creating a word–category co-occurrence matrix. A word–category co-occurrence matrix (WCCM)** is created having word types  $w$  as one dimension and thesaurus categories  $c$  as another.

	$c_1$	$c_2$	...	$c_j$	...
$w_1$	$m_{11}$	$m_{12}$	...	$m_{1j}$	...
$w_2$	$m_{21}$	$m_{22}$	...	$m_{2j}$	...
⋮	⋮	⋮	⋱	⋮	⋮
$w_i$	$m_{i1}$	$m_{i2}$	...	$m_{ij}$	...
⋮	⋮	⋮	...	⋮	⋱

The matrix is populated with co-occurrence counts from a large corpus. A particular cell  $m_{ij}$ , corresponding to word  $w_i$  and category or concept  $c_j$ , is populated with the

number of times  $w_i$  co-occurs (in a window of  $\pm 5$  words) with any word that has  $c_j$  as one of its senses (i.e.,  $w_i$  co-occurs with any word listed under concept  $c_j$  in the thesaurus). For example, assume that the concept of CELESTIAL BODY is represented by four words in the thesaurus: *constellation*, *planet*, *star*, and *sun*. If the word *space* co-occurs with *constellation* (15 times), *planet* (50 times), *star* (40 times), and *sun* (65 times) in the given text corpus, then the cell for *space* and CELESTIAL BODY in the WCCM is populated with 170 (15 + 50 + 40 + 65). This matrix, created after a first pass of the corpus, is called the **base word–category co-occurrence matrix (base WCCM)**.

The choice of  $\pm 5$  words as window size is somewhat arbitrary and hinges on the intuition that words close to a target word are more indicative of its semantic properties than those more distant. Church and Hanks (1990), in their seminal work on word–word co-occurrence association, also use a window size of  $\pm 5$  words and argue that this size is large enough to capture many verb–argument dependencies and yet small enough that adjacency information is not diluted too much.

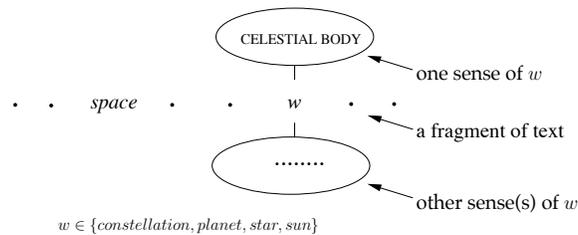
A contingency table for any particular word  $w$  and category  $c$  can be easily generated from the WCCM by collapsing cells for all other words and categories into one and summing up their frequencies.

$$\begin{array}{c|cc}
 & c & \neg c \\
 \hline
 w & n_{wc} & n_{w\neg} \\
 \hline
 \neg w & n_{\neg c} & n_{\neg\neg}
 \end{array}$$

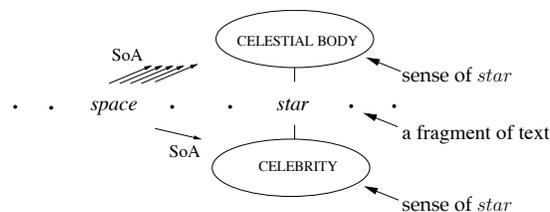
The application of a suitable statistic, such as pointwise mutual information or conditional probability, will then yield the strength of co-occurrence association between the word and the category.

As the base WCCM is created from unannotated text, it will be noisy. For example, out of the 40 times *star* co-occurs with *space*, 25 times it may have been used in the CELESTIAL BODY sense and 15 times in the CELEBRITY sense. However, since this information was not known to the system, the cell for *space*—CELESTIAL BODY in the base WCCM was incremented by 40 rather than 25. Similarly, the cell for *space*—CELEBRITY was also incremented by 40 rather than 15. That said, the base WCCM does capture strong word–category co-occurrence 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. For example, even though we increment counts for both *space*—CELESTIAL BODY and *space*—CELEBRITY for a particular instance where *space* co-occurs with *star*, *space* will co-occur with a number of words such as *planet*, *sun*, and *constellation* that each have the sense of *celestial body* in common (Figure 2), whereas all their other senses are likely different and distributed across the set of concepts. Therefore, the co-occurrence count, and thereby strength of association (SoA), of *space* and CELESTIAL BODY will be relatively higher than that of *space* and CELEBRITY (Figure 3). For more details, see discussion of the general principle by Resnik (1998).

**4.3.2 Bootstrapping.** We now discuss a bootstrapping procedure aimed at reducing, even more, the errors in the WCCM due to word sense ambiguity. Words that occur close to a target word tend to be good indicators of its intended sense. Therefore, a second pass of the corpus is made and the base WCCM is used to roughly disambiguate the words in it. Each word in the corpus is considered as the target one at a time. For each sense of the target, its strength of association with each of the words in its context ( $\pm 5$  words) is summed. The sense that has the highest cumulative association with co-

**Figure 2**

The word *space* will co-occur with a number of words  $X$  that each have one sense of CELESTIAL BODY in common.

**Figure 3**

The base WCCM captures strong word–category co-occurrence strength of association (SoA).

occurring words is chosen as the intended sense of the target word. In this second pass, a new **bootstrapped WCCM** is created such that each cell  $m_{ij}$ , corresponding to word  $w_i$  and concept  $c_j$ , is populated with the number of times  $w_i$  co-occurs with any word *used in sense*  $c_j$ . For example, consider again the 40 times *star* co-occurs with *space*. If the contexts of 25 of these instances have higher cumulative strength of association with CELESTIAL BODY than CELEBRITY, suggesting that in only these 25 of those 40 occurrences *star* was used in CELESTIAL BODY sense, then the cell for *space*–CELESTIAL BODY is incremented by 25 rather than 40 (as was the case in the base WCCM). This bootstrapped WCCM, created after simple and fast word sense disambiguation, will better capture word–concept co-occurrence values, and hence strengths of association values, than the base WCCM.<sup>4</sup>

The bootstrapping step can be repeated; however, further iterations do not improve results significantly. This is not surprising because the base WCCM was created without any word sense disambiguation and so the first bootstrapping iteration with word sense disambiguation will markedly improve the matrix. The same is not true for subsequent iterations.

## 5. Evaluation: monolingual tasks

We evaluated the distributional concept-distance measures on two monolingual tasks: ranking word pairs in order of their semantic distance and correcting real-word spelling errors. Each of these experiments is described in the subsections below. We also used

<sup>4</sup> Speed of disambiguation is important here as all words in the corpus are to be disambiguated. After determining co-occurrence counts from the BNC (a 100 million word corpus), creating the bootstrapped WCCM from the base WCCM took only about 4 hours on a 1.3GHz machine with 16GB memory.

distributional profiles of concepts to determine word sense dominance and obtained near-upper-bound results (not described here; see Mohammad and Hirst (2006a)).

We conducted experiments with four vector distance measures:  $\alpha$ -skew divergence (ASD) ( $\alpha = 0.99$ ), cosine (Cos), Jensen–Shannon divergence (JSD), and Lin’s distributional measure ( $\text{Lin}_{\text{dist}}$ ). All four vector distance measures were used to solve the word-pair ranking and spelling correction tasks in two different ways: (1) by calculating traditional distributional word-distance, that is, distance between distributional profiles of words; (2) by calculating distributional concept-distance, that is, distance between distributional profiles of concepts. This allows for a fair comparison of the two approaches. However, comparison with WordNet-based measures is not so straightforward. Both of the above-mentioned semantic distance tasks have traditionally been performed using WordNet-based measures—which are good at estimating semantic similarity between nouns but particularly poor at estimating semantic relatedness between concept pairs other than noun–noun. This has resulted in the creation of “gold-standard” data only for nouns. As creating new gold-standard data is arduous, we perform experiments on existing noun data.

The distributional profiles of concepts were created from the *British National Corpus* (BNC) and the *Macquarie Thesaurus*. In the base WCCM, 22.85% of the  $98,000 \times 812$  cells had non-zero values whereas the statistic in the bootstrapped WCCM was 9.1%.<sup>5</sup> The word-distance measures used a word–word co-occurrence matrix created from the BNC alone. The BNC was not lemmatized, part-of-speech tagged, nor chunked. The vocabulary was restricted to the words present in the thesaurus (about 98,000 word types) both to provide a level evaluation platform and to filter out named entities and tokens that are not actually words (for example, the BNC has *Hahahahahahahaha*, *perampam*, and *Owzeeyyaaaah*). Also, in order to overcome large computation times of distributional word-distance measures, co-occurrence counts less than five were reset to zero, and words that co-occurred with more than 2000 other words were stoplisted (543 in all). This resulted in a word–word co-occurrence matrix having non-zero values in 0.02% of its  $98,000 \times 98,000$  cells.

### 5.1 Ranking word pairs

A direct approach to evaluate semantic distance measures is to determine how close they are to human judgment and intuition. Given a set of word-pairs, humans can rank them in order of their distance—placing near-synonyms on one end of the ranking and unrelated pairs on the other. As described earlier in Section 2.3, Rubenstein and Goodenough (1965a) provide a “gold-standard” list of 65 human-ranked word-pairs (based on the responses of 51 subjects). An automatic distance measure is deemed to be more accurate than another if its ranking of word-pairs correlates more closely with the human ranking. The concept distance measures were used to determine word-pair distance by two different methods: (1) calculating the concept-distance between all pairs of senses of the two words, and then choosing the shortest distance; (2) taking the average of the distance between each of the relevant pairs of senses.

Table 3 lists correlations of human rankings with those created using the word–word co-occurrence matrix–based traditional distributional word-distance measures and the correlations using the proposed word–concept co-occurrence matrix–based distributional concept-distance measures. Observe that the distributional concept-distance

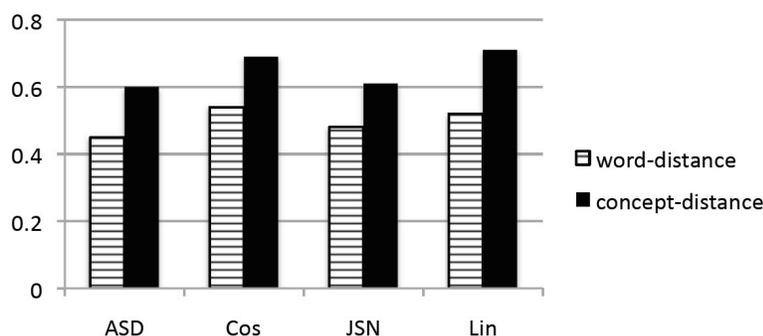
---

<sup>5</sup> Recall that the *Macquarie Thesaurus* has 98,000 word types and 812 categories.

**Table 3**

Correlations with human ranking of Rubenstein and Goodenough word pairs of automatic rankings using traditional word–word co-occurrence–based distributional word-distance measures and the word–concept co-occurrence–based distributional concept-distance measures. Best results for each measure-type are shown in boldface.

Distributional measure	Measure-type		
	Word-distance	Concept-distance	
		closest	average
$\alpha$ -skew divergence	0.45	0.60	–
cosine	<b>0.54</b>	0.69	0.42
Jensen–Shannon divergence	0.48	0.61	–
Lin’s distributional measure	0.52	<b>0.71</b>	<b>0.59</b>

**Figure 4**

Correlations with human ranking of Rubenstein and Goodenough word pairs of automatic rankings using traditional word–word co-occurrence–based distributional word-distance measures and the word–concept co-occurrence–based distributional concept-distance measures.

measures give markedly higher correlation values than distributional word-distance measures. (Figure 4 depicts the results in a graph.) These results are also better than the best reported results (0.64) using latent semantic analysis (Landauer, Foltz, and Laham 1998). Also, using the distance of the closest sense pair (for Cos and Lin<sub>dist</sub>) gives much better results than using the average distance of all relevant sense pairs. (We do not report average distance for ASD and JSD because they give very large distance values when sense-pairs are unrelated; these values dominate the averages, overwhelm the others, and make the results meaningless.) These correlations are, however, notably lower than those obtained by the best WordNet-based measures (not shown in the table), which fall in the range 0.78 to 0.84 (Budanitsky and Hirst 2006).

## 5.2 Correcting real-word spelling errors

The set of Rubenstein and Goodenough word pairs is much too small to safely assume that measures that work well on them do so for the entire English vocabulary. Consequently, semantic measures have traditionally been evaluated through more extensive applications such as the work by Hirst and Budanitsky (2005) on correcting **real-word**

**spelling errors** (or **malapropisms**). If a word in a text is not semantically close to any other word in its context, then it is considered a **suspect**. If the suspect has a spelling-variant that *is* semantically close to a word in its context, then the suspect is declared a probable real-word spelling error and an **alarm** is raised; the semantically close spelling-variant is considered its **correction**. Hirst and Budanitsky tested the method on 500 articles from the 1987–89 *Wall Street Journal* corpus for their experiments, replacing one noun in every 200th word by a spelling-variant and looking at whether the method could restore the original word. This resulted in text with 1408 real-word spelling errors out of a total of 107,233 noun tokens. we adopt this method and this test data, but whereas Hirst and Budanitsky used WordNet-based semantic measures, we use distributional concept- and word-distance measures.

In order to determine whether two words are “semantically close” or not as per any measure of distance, a **threshold** must be set. If the distance between two words is less than the threshold, then they will be considered **semantically close**. Hirst and Budanitsky (2005) pointed out that there is a notably wide band in the human ratings of the Rubenstein and Goodenough word pairs such that no word-pair was assigned a distance value between 1.83 and 2.36 (on a scale of 0–4). They argue that somewhere within this band is a suitable threshold between semantically close and semantically distant, and therefore set thresholds for the WordNet-based measures such that there was maximum overlap in what the automatic measures and human judgments considered semantically close and distant. Following this idea, we use an automatic method to determine thresholds for the various distributional concept- and word-distance measures. Given a list of Rubenstein and Goodenough word pairs ordered according to a distance measure, we repeatedly consider the mean of all adjacent distance values as **candidate thresholds**. Then we determine the number of word-pairs correctly classified as semantically close or semantically distant for each candidate threshold, considering which side of the band they lie as per human judgments. The candidate threshold with highest accuracy is chosen as the threshold.

We follow the Hirst and St. Onge (1998) metrics to evaluate real-word spelling correction. **Suspect ratio** and **alarm ratio** evaluate the processes of identifying suspects and raising alarms, respectively.

$$suspect\ ratio = \frac{(\text{number of true-suspects})/(\text{number of malapropisms})}{(\text{number of false-suspects})/(\text{number of non-malapropisms})} \quad (4)$$

$$alarm\ ratio = \frac{(\text{number of true-alarms})(\text{number of true-suspects})}{(\text{number of false-alarms})/(\text{number of false-suspects})} \quad (5)$$

**Detection ratio** is the product of the two, and measures overall performance in detecting the errors.

$$detection\ ratio = \frac{(\text{number of true-alarms})/(\text{number of malapropisms})}{(\text{number of false-alarms})/(\text{number of non-malapropisms})} \quad (6)$$

**Correction ratio** indicates overall correction performance, and is the “bottom-line” statistic.

$$\text{correction ratio} = \frac{(\text{number of corrected malapropisms})/(\text{number of malapropisms})}{(\text{number of false-alarms})/(\text{number of non-malapropisms})} \quad (7)$$

Values greater than 1 for each of these ratios indicate results better than random guessing. The ability of the system to determine the intended word, given that it has correctly detected an error, is indicated by the **correction accuracy** (0 to 1).

$$\text{correction accuracy} = \frac{\text{number of corrected malapropisms}}{\text{number of true-alarms}} \quad (8)$$

Notice that the correction ratio is the product of the detection ratio and correction accuracy. The overall (single-point) precision (P), recall(R), and F-score (F) of detection are also computed.

$$P = \frac{\text{number of true-alarms}}{\text{number of alarms}} \quad (9)$$

$$R = \frac{\text{number of true-alarms}}{\text{number of malapropisms}} \quad (10)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (11)$$

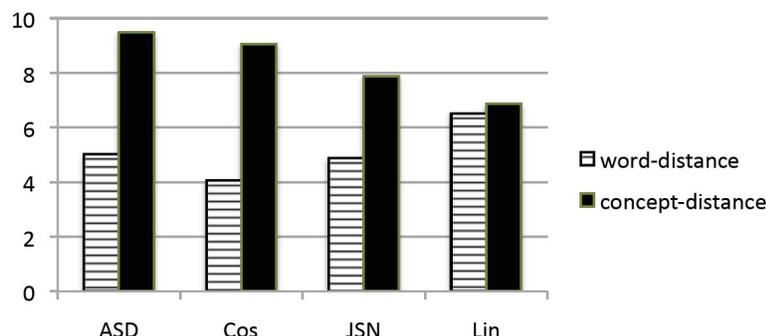
The product of detection F-score and correction accuracy, which we will call **correction performance**, can also be used as a bottom-line performance metric.

Table 4 details the performance of distributional word- and concept-distance measures. For comparison, the table also lists results obtained by Hirst and Budanitsky (2005) using WordNet-based concept-distance measures: those of Hirst and St. Onge (1998), Jiang and Conrath (1997), Leacock and Chodorow (1998), Lin (1997), and Resnik (1995). The last two are information content measures that rely on finding the lowest common subsumer (lcs) of the target synsets in WordNet’s hypernym hierarchy and use corpus counts to determine how specific or general this concept is. The more specific the lcs is and the smaller the difference of its specificity with that of the target concepts, the closer the target concepts are considered. (See Budanitsky and Hirst (2001) for more details.)

Observe that the correction ratio results for the distributional word-distance measures are poor compared to distributional concept-distance measures; the concept-distance measures are clearly superior, in particular  $\alpha$ -skew divergence and cosine. (Figure 5 depicts the results in a graph.) Moreover, if we consider correction ratio to be the bottom-line statistic, then three of the four distributional concept-distance measures outperform all the WordNet-based measures except the Jiang–Conrath measure. If we consider correction performance to be the bottom-line statistic, then again we see that the distributional concept-distance measures outperform the word-distance measures, except in the case of Lin’s distributional measure, which gives slightly poorer results with concept-distance.

**Table 4** Real-word spelling error correction. The best results as per the two bottom-line statistics, correction ratio and correction performance, are shown in boldface.

Measure	suspect ratio	alarm ratio	detection ratio	correction accuracy	correction ratio	p detection	K detection	F	correction performance
<i>Distributional word</i>									
$\alpha$ -skew divergence	3.36	1.78	5.98	0.84	5.03	7.37	45.53	12.69	10.66
cosine	2.91	1.64	4.77	0.85	4.06	5.97	37.15	10.28	8.74
Jensen-Shannon divergence	3.29	1.77	5.82	0.83	4.88	7.19	44.32	12.37	10.27
<b>Lin's distributional measure</b>	3.63	2.15	7.78	0.84	<b>6.52</b>	9.38	58.38	16.16	<b>13.57</b>
<i>Distributional concept</i>									
$\alpha$ -skew divergence	4.11	2.54	10.43	0.91	<b>9.49</b>	12.19	25.28	16.44	<b>14.96</b>
cosine	4.00	2.51	10.03	0.90	9.05	11.77	26.99	16.38	14.74
Jensen-Shannon divergence	3.58	2.46	8.79	0.90	7.87	10.47	34.66	16.08	14.47
Lin's distributional measure	3.02	2.60	7.84	0.88	6.87	9.45	36.86	15.04	13.24
<i>WNNet concept</i>									
Hirst-St-Onge	4.24	1.95	8.27	0.93	7.70	9.67	26.33	14.15	13.16
<b>Jiang-Conrath</b>	4.73	2.97	14.02	0.92	<b>12.91</b>	14.33	46.22	21.88	<b>20.13</b>
Leacock-Chodrow	3.23	2.72	8.80	0.83	7.30	11.56	60.33	19.40	16.10
Lin's WordNet-based measure	3.57	2.71	9.70	0.87	8.48	9.56	51.56	16.13	14.03
Resnik	2.58	2.75	7.10	0.78	5.55	9.00	55.00	15.47	12.07



**Figure 5**

Correction ratio obtained on the real-word spelling correction task using traditional word–word co-occurrence–based distributional word-distance measures and the word–concept co-occurrence–based distributional concept-distance measures.

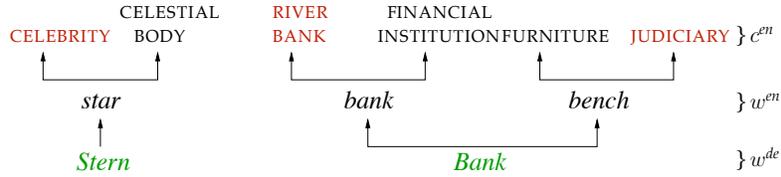
Also, in contrast to correction ratio values, using the Leacock–Chodorow measure results in relatively higher correction performance values than the best distributional concept-distance measures. While it is clear that the Leacock–Chodorow measure is relatively less accurate in choosing the right spelling-variant for an alarm (correction accuracy), detection ratio and detection  $F$ -score present contrary pictures of relative performance in detection.

As the correction ratio is determined by the product of a number of ratios, each evaluating the various stages of malapropism correction (identifying suspects, raising alarms, and applying the correction), we believe it is a better indicator of overall performance than correction performance, which is a not-so-elegant product of an  $F$ -score and accuracy. However, no matter which of the two is chosen as the bottom-line performance statistic, the results show that the distributional concept-distance measures are clearly superior to word-distance measures. Further, of all the WordNet-based measures, only that proposed by Jiang and Conrath outperforms the best distributional concept-distance measures consistently with respect to both bottom-line statistics.

## 6. Cross-lingual Semantic Distance

The application of semantic distance algorithms in most languages is hindered by the lack of high-quality linguistic resources. WordNet-based measures of semantic distance, such as those of Jiang and Conrath (1997) and Resnik (1995), require a WordNet. Distributional measures of word-distance, as shown in Section 4 earlier, are markedly less accurate because they conflate the many senses of a word. Also as shown there, distributional measures of concept-distance avoid sense conflation and achieve results better than the traditional word-distance measures. However, the high-quality thesauri and WordNet-like resources that the concept-distance methods require do not exist for most of the 3000–6000 languages in existence today and they are costly to create.

Here we show how distributional measures of concept-distance can be ported to a cross-lingual framework to overcome this knowledge bottleneck. We describe **cross-lingual distributional measures of concept-distance**, or simply **cross-lingual measures**, that determine the distance between a word pair in resource-poor language  $L_1$  using a knowledge source in a resource-rich language  $L_2$ . We use a bilingual lexicon to connect the words in  $L_1$  with the words in  $L_2$ . We will compare this approach with the


**Figure 6**

The cross-lingual candidate senses of German words *Stern* and *Bank*. In red are concepts that are not really senses of the German words, but simply artifacts of the translation step.

best monolingual approaches, which usually require high-quality knowledge sources in the same language ( $L_1$ ); the smaller the loss in performance, the more capable the cross-lingual algorithm is of overcoming ambiguities in word translation. An evaluation, therefore, requires an  $L_1$  that in actuality has adequate knowledge sources. Therefore we chose German to stand in as the resource-poor language  $L_1$  and English as the resource-rich  $L_2$ . The evaluation tasks will involve estimating the semantic distance between German words. Both monolingual and cross-lingual approaches will use the same German corpus, but while the monolingual approach will use a knowledge source in the same language, the German GermaNet, the cross-lingual approach (which we describe below) will use a knowledge source from another language, the English *Macquarie Thesaurus*. The remainder of this section describes our approach in terms of German and English, but the algorithm itself is language independent.

### 6.1 Cross-lingual senses, cross-lingual distributional profiles, and cross-lingual distributional distance

Given a German word  $w^{de}$  in context, we use a German–English bilingual lexicon to determine its different possible English translations. Each English translation  $w^{en}$  may have one or more possible coarse senses, as listed in an English thesaurus. These English thesaurus concepts ( $c^{en}$ ) will be referred to as the **cross-lingual candidate senses** of the German word  $w^{de}$ . Figure 6 depicts examples. They are called “candidate” because some of the senses of  $w^{en}$  might not really be senses of  $w^{de}$ . For example, CELESTIAL BODY and CELEBRITY are both senses of the English word *star*, but the German word *Stern* can mean only CELESTIAL BODY and not CELEBRITY. Similarly, the German *Bank* can mean FINANCIAL INSTITUTION or FURNITURE, but not RIVER BANK or JUDICIARY. An automated system has no straightforward method of teasing out the actual cross-lingual senses of  $w^{de}$  from those that are an artifact of the translation step. So we treat them all as its senses. Now, we proceed to determine semantic distance just as in the monolingual case, except that the words are German and their senses are English thesaurus categories. Table 5 presents a mini vocabulary of German words needed to understand the discussion below.

As in the monolingual estimation of distributional concept-distance, the distance between two concepts is calculated by first determining their DPs (co-occurrence vectors). Recall the example monolingual DPs of the two senses of *star*:

CELESTIAL BODY (*celestial body, sun, ...*): *space* 0.36, *light* 0.27, *constellation* 0.11, *hydrogen* 0.07, ...

CELEBRITY (*celebrity, hero, ...*): *famous* 0.24, *movie* 0.14, *rich* 0.14, *fan* 0.10, ...

**Table 5**

Vocabulary of German words needed to understand this discussion.

German word	Meaning(s)	German word	Meaning(s)
<i>Bank</i>	1. financial institution 2. bench (furniture)	<i>Licht</i>	light
<i>berühmt</i>	famous	<i>Morgensonne</i>	morning sun
<i>Bombe</i>	bomb	<i>Raum</i>	space
<i>Erwärmung</i>	heat	<i>reich</i>	rich
<i>Film</i>	movie (motion picture)	<i>Sonne</i>	sun
<i>Himmelskörper</i>	heavenly body	<i>Star</i>	star (celebrity)
<i>Konstellation</i>	constellation	<i>Stern</i>	star (celestial body)
		<i>Verschmelzung</i>	fusion

In the cross-lingual approach, a concept is now glossed by near-synonymous words in an *English* thesaurus, whereas its profile is made up of the strengths of association with co-occurring *German* words. We will call them **cross-lingual distributional profiles of concepts** or just **cross-lingual DPCs**. Here are constructed examples for the two cross-lingual candidate senses of the German word *Stern*:

CELESTIAL BODY (*celestial body, sun, ...*): *Raum* 0.36, *Licht* 0.27, *Konstellation* 0.11, ...  
 CELEBRITY (*celebrity, hero, ...*): *berühmt* 0.24, *Film* 0.14, *reich* 0.14, ...

The values are the strength of association (usually pointwise mutual information or conditional probability) of the target concept with co-occurring words. In order to calculate the strength of association, we must first determine individual word and concept counts, as well as their co-occurrence counts. The next section describes how these can be estimated without the use of any word-aligned parallel corpora and without any sense-annotated data. The closer the cross-lingual DPs of two concepts, the smaller is their semantic distance. Just as in the case of monolingual distributional concept-distance measures (described in Section 4.2 earlier), distributional measures can be used to estimate the distance between the cross-lingual DPs of two target concepts. For example, recall how cosine is used in a monolingual framework to estimate distributional distance between two concepts:

$$\text{Cos}_{cp}(c_1, c_2) = \frac{\sum_{w \in C(c_1) \cup C(c_2)} (P(w|c_1) \times P(w|c_2))}{\sqrt{\sum_{w \in C(c_1)} P(w|c_1)^2} \times \sqrt{\sum_{w \in C(c_2)} P(w|c_2)^2}} \quad (12)$$

$C(x)$  is the set of English words that co-occur with English *concept*  $x$  within a pre-determined window. The conditional probabilities in the formula are taken from the monolingual distributional profiles of concepts. We can adapt the formula to estimate cross-lingual distributional distance between two concepts as shown below:

$$\text{Cos}_{cp}(c_1^{en}, c_2^{en}) = \frac{\sum_{w^{de} \in C(c_1^{en}) \cup C(c_2^{en})} (P(w^{de}|c_1^{en}) \times P(w^{de}|c_2^{en}))}{\sqrt{\sum_{w^{de} \in C(c_1^{en})} P(w^{de}|c_1^{en})^2} \times \sqrt{\sum_{w^{de} \in C(c_2^{en})} P(w^{de}|c_2^{en})^2}} \quad (13)$$

$C(x)$  is now the set of German words that co-occur with English concept  $x$  within a pre-determined window. The conditional probabilities in the formula are taken from the cross-lingual DPCs.

If the distance between two German words is required, then the distance between all relevant English cross-lingual candidate sense pairs is determined and the minimum is chosen. For example, if *Stern* has the two cross-lingual candidate senses mentioned above and *Verschmelzung* has one (FUSION), then the distance between them is determined by first applying cosine (or any distributional measure) to the cross-lingual DPs of CELESTIAL BODY and FUSION:

CELESTIAL BODY (*celestial body, sun, ...*): Raum 0.36, Licht 0.27, Konstellation 0.11, ...  
 FUSION (*thermonuclear reaction, atomic reaction, ...*): Erwärmung 0.16, Bombe 0.09, Licht 0.09, Raum 0.04, ...

Then applying cosine to the cross-lingual DPs of CELEBRITY and FUSION:

CELEBRITY (*celebrity, hero, ...*): berühmt 0.24, Film 0.14, reich 0.14, ...  
 FUSION (*thermonuclear reaction, atomic reaction, ...*): Erwärmung 0.16, Bombe 0.09, Licht 0.09, Raum 0.04, ...

And finally choosing the one with minimum semantic distance, that is, maximum similarity/relatedness:

$$\text{distance}(\textit{Stern}, \textit{Verschmelzung}) = \max(\text{Cos}_{cp}(\text{CELEBRITY}, \text{FUSION}), \text{Cos}_{cp}(\text{CELESTIAL BODY}, \text{FUSION})) \quad (14)$$

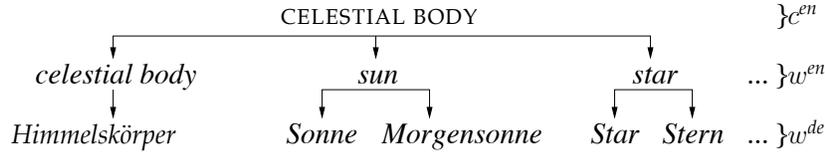
Maximum is chosen because cosine is a similarity/relatedness measure. In case of distance measures, such as  $\alpha$  Skew Divergence, the minimum will be chosen.

## 6.2 Estimating cross-lingual DPCs by creating cross-lingual word–category co-occurrence matrix

Determining cross-lingual distributional profiles of concepts requires information about which words in one language  $L_1$  co-occur with which concepts as defined in another language  $L_2$ . This means that a direct approach requires the text in  $L_1$ , from which counts are made, to have a word-aligned parallel corpus in  $L_2$ . Further, the  $L_2$  text must be sense annotated. Such data exists rarely, if at all, and it is expensive to create. Here we present a way to estimate cross-lingual distributional profiles of concepts from raw text (in one language,  $L_1$ ) and a published thesaurus (in another language,  $L_2$ ) using an  $L_1$ – $L_2$  bilingual lexicon and a bootstrapping algorithm.

We create a cross-lingual word–category co-occurrence matrix with German word types  $w^{de}$  as one dimension and English thesaurus concepts  $c^{en}$  as the other.

	$c_1^{en}$	$c_2^{en}$	...	$c_j^{en}$	...
$w_1^{de}$	$m_{11}$	$m_{12}$	...	$m_{1j}$	...
$w_2^{de}$	$m_{21}$	$m_{22}$	...	$m_{2j}$	...
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$w_i^{de}$	$m_{i1}$	$m_{i2}$	...	$m_{ij}$	...
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	$\ddots$



**Figure 7**  
Words having CELESTIAL BODY as one of their cross-lingual candidate senses.

The matrix is populated with co-occurrence counts from a large German corpus.

A particular cell  $m_{ij}$ , corresponding to word  $w_i^{de}$  and concept  $c_j^{en}$ , is populated with the number of times the German word  $w_i^{de}$  co-occurs (in a window of  $\pm 5$  words) with any German word having  $c_j^{en}$  as one of its *cross-lingual candidate senses*. For example, the *Raum*–CELESTIAL BODY cell will have the sum of the number of times *Raum* co-occurs with *Himmelskörper*, *Sonne*, *Morgensonne*, *Star*, *Stern*, and so on (see Figure 7). This matrix, created after a first pass of the corpus, is called the **cross-lingual base WCCM**. A contingency table for any particular German word  $w^{de}$  and English category  $c^{en}$  can be easily generated from the WCCM by collapsing cells for all other words and categories into one and summing up their frequencies.

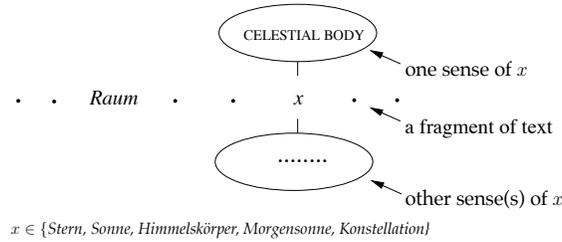
$$\begin{array}{c|cc}
 & c^{en} & \neg c^{en} \\
 \hline
 w^{de} & n_{w^{de}c^{en}} & n_{w^{de}\neg} \\
 \neg w^{de} & n_{\neg c^{en}} & n_{\neg}
 \end{array}$$

The application of a suitable statistic, such as PMI or conditional probability, will then yield the strength of association between the German word and the English category.

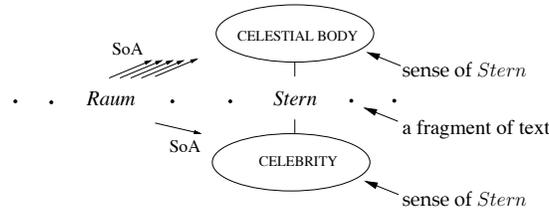
As the cross-lingual base WCCM is created from unannotated text, it will be noisy (for the same word-sense-ambiguity reasons as to why the monolingual base WCCM is noisy—explained in Section 4.3.1 earlier). Yet, again, the cross-lingual base WCCM does capture strong associations between a category (concept) and co-occurring words (just like the monolingual base WCCM). For example, even though we increment counts for both *Raum*–CELESTIAL BODY and *Raum*–CELEBRITY for a particular instance where *Raum* co-occurs with *Star*, *Raum* will co-occur with a number of words such as *Himmelskörper*, *Sonne*, and *Morgensonne* that each have the sense of CELESTIAL BODY in common (see Figures 7 and 8), whereas all their other senses are likely different and distributed across the set of concepts. Therefore, the co-occurrence count of *Raum* and CELESTIAL BODY, and thereby their strength of association, will be relatively higher than those of *Raum* and CELEBRITY (Figure 9). Therefore, we bootstrap the matrix just as described before in the monolingual case (Section 4.3.2).

### 7. Evaluation: cross-lingual, word-distance tasks

We evaluated German–English cross-lingual distributional measures of concept-distance on the tasks of (1) measuring semantic distance between German words and ranking German word pairs according to semantic distance, and (2) solving German ‘nearest-synonym’ questions from *Reader’s Digest*. Each of these is described in the subsections below. We also used Chinese–English cross-lingual distributional profiles of concepts in SemEval-2007’s Chinese–English word-translation task (not described here; see Mohammad et al. (2007)).



**Figure 8**  
The word *Raum* will also co-occur with a number of other words *x* that each have one sense of CELESTIAL BODY in common.



**Figure 9**  
The base WCCM captures strong word–category co-occurrence strength of association (SoA).

The German-English distributional profiles were created using the following resources: the German newspaper corpus *taz*<sup>6</sup> (Sep 1986 to May 1999; 240 million words), the English *Macquarie Thesaurus* (Bernard 1986) (about 98,000 word types), and the German-English bilingual lexicon BEOLINGUS<sup>7</sup> (about 265,000 entries). Multi-word expressions in the thesaurus and the bilingual lexicon were ignored. We used a context of  $\pm 5$  words on either side of the target word for creating the base and bootstrapped WCCMs. No syntactic pre-processing was done, nor were the words stemmed, lemmatized, or part-of-speech tagged.

In order to compare results with state-of-the-art monolingual approaches we conducted experiments using GermaNet measures as well. The specific distributional measures and GermaNet-based measures used are listed in Table 6. The GermaNet measures used are of two kinds: (1) information content measures, and (2) Lesk-like measures that rely on *n*-gram overlaps in the glosses of the target senses, proposed by Gurevych (2005). As GermaNet does not have glosses for synsets, Gurevych (2005) proposed a way of creating a bag-of-words-type pseudo-gloss for a synset by including the words in the synset and in synsets close to it in the network. The information content measures rely on finding the lowest common subsumer (lcs) of the target synsets in a hypernym hierarchy and using corpus counts to determine how specific or general this concept is. The more specific the lcs is and the smaller the difference of its specificity with that of the target concepts, the closer the target concepts are.

<sup>6</sup> <http://www.taz.de>  
<sup>7</sup> <http://dict.tu-chemnitz.de>

**Table 6**

Distance measures used in the experiments.

(Cross-lingual) Distributional Measures	(Monolingual) GermaNet Measures	
	Information Content-based	Lesk-like
$\alpha$ -skew divergence (Lee 2001)	Jiang and Conrath (1997)	hypernym pseudo-gloss
cosine (Schütze and Pedersen 1997)	Lin (1998b)	(Gurevych 2005)
Jensen-Shannon divergence (Dagan, Lee, and Pereira 1994)	Resnik (1995)	radial pseudo-gloss
Lin (1998a)		(Gurevych 2005)

## 7.1 Ranking word pairs

**7.1.1 Data.** Gurevych (2005) and Zesch et al. (2007) asked native German speakers to mark two different sets of German word pairs with distance values. Set 1 (**Gur65**) is the German translation of the English (Rubenstein and Goodenough 1965b) dataset. It has 65 noun–noun word pairs. Set 2 (**Gur350**) is a larger dataset containing 350 word pairs made up of nouns, verbs, and adjectives. The semantically close word pairs in Gur65 are mostly synonyms or hypernyms (hyponyms) of each other, whereas those in Gur350 have both classical and non-classical relations (Morris and Hirst 2004) with each other. Details of these **semantic distance benchmarks**<sup>8</sup> were listed earlier in Table 1.

**7.1.2 Results and Discussion.** Word-pair distances determined using different distance measures are compared in two ways with the two human-created benchmarks. The rank ordering of the pairs from closest to most distant is evaluated with Spearman’s rank order correlation  $\rho$ ; the distance judgments themselves are evaluated with Pearson’s correlation coefficient  $r$ . The higher the correlation, the more accurate the measure is. Spearman’s correlation ignores actual distance values after a list is ranked—only the ranks of the two sets of word pairs are compared to determine correlation. On the other hand, Pearson’s coefficient takes into account actual distance values. So even if two lists are ranked the same, but one has distances between consecutively-ranked word-pairs more in line with human-annotations of distance than the other, then Pearson’s coefficient will capture this difference. However, this makes Pearson’s coefficient sensitive to outlier data points, and so one must interpret it with caution. Therefore, Spearman’s rank correlation is more common in the semantic distance literature. However, many of the experiments on German data report Pearson’s correlation. We report both correlations in Table 7.

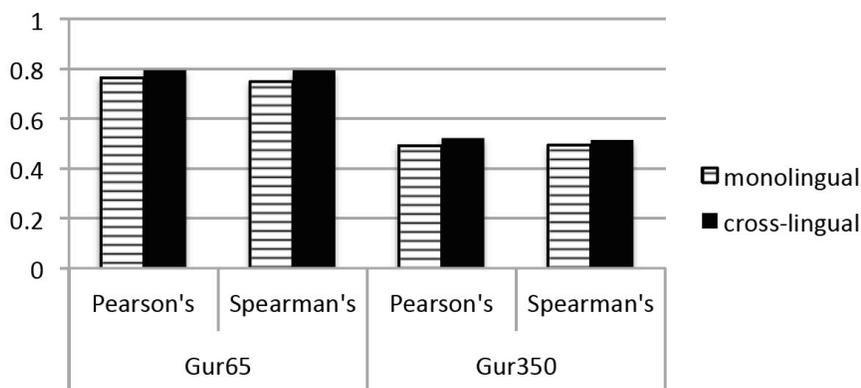
Observe that on both datasets and by both measures of correlation, the cross-lingual measures of concept-distance perform not just as well as the best monolingual measures, but in fact slightly better. (Figure 10 depicts the results in a graph.) In general, the correlations are lower for Gur350 as it contains cross-PoS word pairs and non-classical relations, making it harder to judge even by humans (as shown by the inter-annotator correlations for the datasets in Table 1). As per Spearman’s rank correlation,  $\alpha$ -skew divergence and Jensen-Shannon divergence perform best on both datasets. The correlations of cosine and Lin’s distributional measure are not far behind. Amongst the monolingual GermaNet measures, radial pseudo-gloss performs best. As per Pearson’s

<sup>8</sup> The datasets are publicly available at <http://www.ukp.tu-darmstadt.de/data/semRelDatasets>.

**Table 7**

Ranking German word pairs: Correlations of distance measures with human judgments. The best results obtained using monolingual and cross-lingual measures are marked in bold.

Measure	Gur65		Gur350	
	Spearman's rank correlation	Pearson's correlation	Spearman's rank correlation	Pearson's correlation
<i>Monolingual</i>				
hypernym pseudo-gloss	0.672	0.702	0.346	0.331
radial pseudo-gloss	<b>0.764</b>	0.565	<b>0.492</b>	0.420
Jiang and Conrath measure	0.665	<b>0.748</b>	0.417	0.410
Lin's GermaNet measure	0.607	0.739	0.475	<b>0.495</b>
Resnik's measure	0.623	0.722	0.454	0.466
<i>Cross-lingual</i>				
$\alpha$ -skew divergence	<b>0.794</b>	0.597	<b>0.520</b>	0.413
cosine	0.778	0.569	0.500	0.212
Jensen-Shannon divergence	<b>0.793</b>	0.633	<b>0.522</b>	0.422
Lin's distributional measure	0.775	<b>0.816</b>	0.498	<b>0.514</b>


**Figure 10**

Ranking German word pairs: Spearman's rank correlation obtained when using the best cross-lingual distributional concept-distance measure and that obtained when using the best monolingual GermaNet-based measure.

correlation, Lin's distributional measure performs best overall and radial pseudo-gloss does best amongst the monolingual measures.

## 7.2 Solving word choice problems from *Reader's Digest*

**7.2.1 Data.** Our next approach to evaluating distance measures follows that of Jarmasz and Szpakowicz (2003), who evaluated semantic similarity measures through their ability to solve synonym problems (80 TOEFL (Landauer and Dumais 1997), 50 ESL (Turney 2001), and 300 (English) *Reader's Digest* Word Power questions). Turney (2006) used a similar approach to evaluate the identification of semantic relations, with 374 college-level multiple-choice word analogy questions.

Issues of the German edition of *Reader's Digest* include a word choice quiz called 'Word Power'. Each question has one target word and four alternative words or phrases;

the objective is to pick the alternative that is most closely related to the target. For example:<sup>9</sup>

- |                                     |                                    |
|-------------------------------------|------------------------------------|
| <i>Duplikat</i> (duplicate)         |                                    |
| a. <i>Einzelstück</i> (single copy) | b. <i>Doppelkinn</i> (double chin) |
| c. <i>Nachbildung</i> (replica)     | d. <i>Zweitschrift</i> (copy)      |

Torsten Zesch compiled the **Reader’s Digest Word Power (RDWP) benchmark** for German, which consists of 1072 of these word-choice problems collected from the January 2001 to December 2005 issues of the German-language edition (Wallace and Wallace 2005). Forty-four problems that had more than one correct answer and twenty problems that used a phrase instead of a single term as the target were discarded. The remaining 1008 problems form our evaluation dataset, which is significantly larger than any of the previous datasets employed in a similar evaluation.

We evaluate the various cross-lingual and monolingual distance measures by their ability to choose the correct answer. The distance between the target and each of the alternatives is computed by a measure, and the alternative that is closest is chosen. If two or more alternatives are equally close to the target, then the alternatives are said to be **tied**. If one of the tied alternatives is the correct answer, then the problem is counted as correctly solved, but the corresponding score is reduced. The system assigns a score of 0.5, 0.33, and 0.25 for 2, 3, and 4 tied alternatives, respectively (in effect approximating the score obtained by randomly guessing one of the tied alternatives). If more than one alternative has a sense in common with the target, then the thesaurus-based cross-lingual measures will mark them each as the closest sense. However, if one or more of these tied alternatives is in the same semicolon group of the thesaurus as the target, then only these are chosen as the closest senses.<sup>10</sup>

Even though we discard questions from the German RDWP dataset that contained a phrasal target, we did not discard questions that had phrasal alternatives simply because of the large number of such questions. Many of these phrases cannot be found in the knowledge sources (GermaNet or *Macquarie Thesaurus* via translation list). In these cases, we remove stopwords (prepositions, articles, etc.) and split the phrase into component words. As German words in a phrase can be highly inflected, all components are lemmatized. For example, the target *imaginär* (*imaginary*) has *nur in der Vorstellung vorhanden* (*exists only in the imagination*) as one of its alternatives. The phrase is split into its component words *nur*, *Vorstellung*, and *vorhanden*. The system computes semantic distance between the target and each phrasal component and selects the minimum value as the distance between target and potential answer.

**7.2.2 Results and Discussion.** Table 8 presents the results obtained on the German RDWP benchmark for both monolingual and cross-lingual measures. Only those questions for which the measures have some distance information are attempted; the column ‘# attempted’ shows the number of questions attempted by each measure, which is the maximum score that the measure can hope to get. Observe that the thesaurus-based cross-lingual measures have a much larger coverage than the GermaNet-based monolingual measures. The cross-lingual measures have a much larger number of correct

<sup>9</sup> English translations are in parentheses.

<sup>10</sup> Words in a thesaurus category are further partitioned into different paragraphs and each paragraph into semicolon groups. Words within a semicolon group are more closely related than those in semicolon groups of the same paragraph or category.

**Table 8**

Solving word choice questions: Performance of monolingual and cross-lingual distance measures. The best results for each class of measures are marked in bold.

Measure	Reader's Digest Word Power benchmark						
	# attempted	# correct	# ties	Score	P	R	F
<i>Monolingual</i>							
hypernym pseudo-gloss	222	174	11	<b>171.5</b>	.77	.17	<b>.28</b>
radial pseudo-gloss	266	188	15	<b>184.7</b>	.69	.18	<b>.29</b>
Jiang and Conrath	357	157	1	156.0	.44	.16	.23
Lin's GermaNet measure	298	153	1	152.5	.51	.15	.23
Resnik's measure	299	154	33	148.3	.50	.15	.23
<i>Cross-lingual</i>							
$\alpha$ -skew divergence	438	185	81	151.6	.35	.15	.21
cosine	438	276	90	<b>223.1</b>	.51	.22	<b>.31</b>
Jensen-Shannon divergence	438	276	90	<b>229.6</b>	.52	.23	<b>.32</b>
Lin's distributional measure	438	274	90	<b>228.7</b>	.52	.23	<b>.32</b>

answers too (column '# correct'), but this number is bloated due to the large number of ties. We see more ties when using the cross-lingual measures because they rely on the *Macquarie Thesaurus*, a very coarse-grained sense inventory (around 800 categories), whereas the monolingual measures operate on the fine-grained GermaNet. 'Score' is the score each measure gets after it is penalized for the ties. The cross-lingual measures cosine, Jensen-Shannon divergence, and Lin's distributional measure obtain the highest scores. But 'Score' by itself does not present the complete picture either as, given the scoring scheme, a measure that attempts more questions may get a higher score just from random guessing. We therefore present precision (P), recall (R), and *F* measure (F):

$$P = \frac{\text{Score}}{\# \text{ attempted}} \quad (15)$$

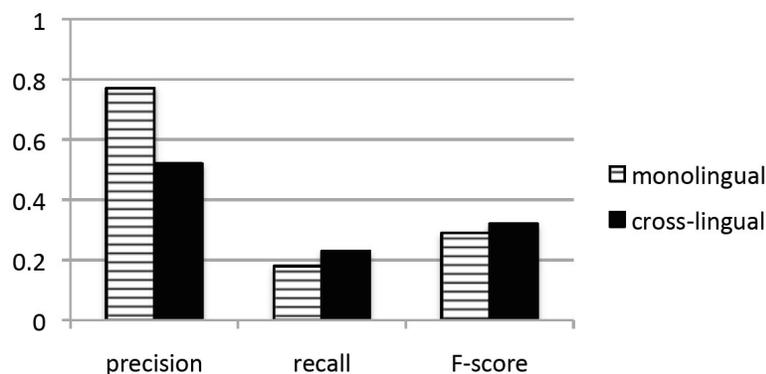
$$R = \frac{\text{Score}}{1008} \quad (16)$$

$$F = \frac{2 \times P \times R}{P + R} \quad (17)$$

Figure 11 depicts the results in a graph. Observe that the cross-lingual measures have a higher coverage (recall) than the monolingual measures but lower precision. The *F* measures show that the best cross-lingual measures do slightly better than the best monolingual ones, despite the large number of ties. The measures of cosine, Jensen-Shannon divergence, and Lin's distributional measure remain the best cross-lingual measures, whereas hypernym pseudo-gloss and radial pseudo-gloss are the best monolingual ones.

## 8. Related work

A detailed survey of WordNet-based semantic distance measures is given by Budanitsky and Hirst (2006). Patwardhan et al. (2003) also compare the performance of various



**Figure 11**  
Solving word choice questions: Performance of the best monolingual and cross-lingual distance measures.

WordNet-based measures. See Curran (2004), Weeds et al. (2004), and Mohammad and Hirst (2007) for comprehensive surveys of distributional measures of word-distance.

Yarowsky (1992) proposed a model for unsupervised word sense disambiguation using *Roget's Thesaurus*. A mutual information-like measure was used to identify words that best represent each category in the thesaurus, which he calls the **salient words**. The presence of a salient word in the context of a target word is evidence that the word is used in a sense corresponding to the salient word. The evidence is incorporated in a Bayesian model. The word-category co-occurrence matrix (WCCM) we created can be seen as a means of determining the degree of salience of any word co-occurring with a concept. We further improved the accuracy of the WCCM using a bootstrapping technique.

Jarmasz and Szpakowicz (2003) use the taxonomic structure of the *Roget's Thesaurus* to determine semantic similarity. Two words are considered maximally similar if they occur in the same semicolon group in the thesaurus. Then on, decreasing in similarity are word pairs in the same paragraph, words pairs in different paragraphs belonging to the same part of speech and within the same category, word pairs in the category, and so on until word pairs which have nothing in common except that they are in the thesaurus (maximally distant). However, a large number of words that are in different thesaurus categories may be semantically related. Thus, this approach is better suited for estimating semantic similarity than semantic relatedness. Our approach is specifically intended to determine the semantic relatedness between word pairs across thesaurus categories.

Pantel and Lin (2002) proposed a method to discover word senses from text using word co-occurrence information. The approach produces clusters of words that are semantically similar and there is a numeric score representing the distance of each word in a cluster with the centroid of that cluster. Note that these clusters do not have information of which words co-occur with the clusters (concepts) and so these are not distributional profiles of concepts (DPCs). Rather, the output of the Pantel and Lin system is more like a *Roget's* or *Macquaries Thesaurus*, except that it is automatically generated. One can create DPCs using our method and the Pantel and Lin thesaurus (instead of *Macquarie*) and it will be interesting to determine its usefulness. However, we suspect that there will be more complementarity between information encoded in a human created lexical resource and the co-occurrence information in text.

Pantel (2005) also provides a way to create co-occurrence vectors for WordNet senses. The lexical co-occurrence vectors of words in a leaf node are propagated up the WordNet hierarchy. A parent node inherits those co-occurrences that are shared by its children. Lastly, co-occurrences not pertaining to the leaf nodes are removed from its vector. Even though the methodology attempts to associate a WordNet node or sense with only those co-occurrences that pertain to it, no attempt is made at correcting the frequency counts. After all, *word1-word2* co-occurrence frequency (or association) is likely not the same as *SENSE1-word2* co-occurrence frequency (or association), simply because *word1* may have senses other than *SENSE1*, as well. Further, in Pantel's system, the co-occurrence frequency associated with a parent node is the weighted sum of co-occurrence frequencies of its children. The frequencies of the child nodes are used as weights. Sense ambiguity issues apart, this is still problematic because a parent concept (say, *BIRD*) may co-occur much more frequently (or infrequently) with a word than its children do. In contrast, the bootstrapped WCCM not only identifies which words co-occur with which concepts, but also has more accurate estimates of the co-occurrence frequencies.

Patwardhan and Pedersen (2006) create **aggregate co-occurrence vectors** for a WordNet sense by adding the co-occurrence vectors of the words in its WordNet gloss. The distance between two senses is then determined by the cosine of the angle between their aggregate vectors. However, such aggregate co-occurrence vectors are expected to be noisy because they are created from data that is not sense-annotated. The bootstrapping procedure introduced in Section 4.3.2 minimizes such errors and, as we showed in Mohammad and Hirst (2006a), markedly improves accuracies of natural language tasks that use these co-occurrence vectors.

Véronis (2004) presents a graph theory-based approach to identify the various senses of a word in a text corpus without the use of a dictionary. For each target word, a graph of inter-connected nodes is created. Every word that co-occurs with the target word is a node. Two nodes are connected with an edge if they are found to co-occur with each other. 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 to quantify evidence that the target word is used in the corresponding sense. However, these strengths of association 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.

Erk and Padó (2008) proposed a way of determining the distributional profile of a word in context. They use dependency relations and selectional preferences of the target words and combine multiple co-occurrence vectors in a manner so as to give more weight to co-occurring words pertaining to the intended senses of the target words. This approach effectively assumes that each occurrence of a word in a different context has a unique meaning. In contrast, our approach explores the use of only about a thousand very coarse concepts to represent the meaning of all words in the vocabulary. By choosing to work with much coarser concepts, the approach foregoes the ability to make fine-grained distinctions in meaning, but is able to better estimate semantic distance between the coarser concepts as there is much more information pertaining to them.

## 9. Conclusion

We proposed an approach that allows distributional measures to estimate semantic distance between *concepts* using a published thesaurus and raw text. Additionally, we showed how this approach can be ported into a cross-lingual framework to determine semantic distance in a resource-poor language by combining its text with a knowledge source in a different, preferably resource-rich, language. We evaluated the approach both monolingually and cross-lingually in comparison with traditional word-distance measures and WordNet-like resource-based concept-distance measures.

The monolingual evaluation required the automatic measures to (1) rank word-pairs in order of their human-judged linguistic distance, and (2) correct real-word spelling errors. The distributional concept-distance measures outperformed word-distance measures in both tasks. They do not perform as well as the best WordNet-based measures in ranking a small set of word pairs, but in the task of correcting real-word spelling errors, they beat all WordNet-based measures except for Jiang–Conrath (which is markedly better) and Leacock–Chodorow (which is slightly better if we consider correction performance as the bottom-line statistic, but slightly worse if we rely on correction ratio). It should be noted that the Rubenstein and Goodenough word-pairs used in the ranking task, as well as all the real-word spelling errors in the correction task, are nouns. We expect that the WordNet-based measures will perform poorly when other parts of speech are involved, as those hierarchies of WordNet are not as extensively developed. Further, the various hierarchies are not well connected, nor is it clear how to use these interconnections across parts of speech for calculating semantic distance. On the other hand, the distributional concept-distance measures do not rely on any hierarchies (even if they exist in a thesaurus) but on sets of words that unambiguously represent each sense. Further, because our measures are tied closely to the corpus from which co-occurrence counts are made, we expect the use of domain-specific corpora to give even better results.

We evaluated the cross-lingual measures against the best monolingual ones operating on a WordNet-like resource, GermaNet, through an extensive set of experiments on two different tasks: (1) rank German word-pairs in order of their human-judged linguistic distance, and (2) solve German ‘Word Power’ questions from *Reader’s Digest*. Even with the added ambiguity of translating words from one language to another, the cross-lingual measures performed slightly better than the best monolingual measures on both the word-pair task and the *Reader’s Digest* word-choice task. Further, in the word-choice task, the cross-lingual measures obtained a significantly higher coverage than the monolingual measure. The richness of English resources seems to have a major impact, even though German, with GermaNet, a well-established resource, is in a better position than most other languages. This is indeed promising, because obtaining broad coverage for resource-poor languages remains an important goal as we integrate state-of-the-art approaches in natural language processing into real-life applications. These results show that the proposed algorithm can successfully combine German text with an English thesaurus using a bilingual German–English lexicon to obtain state-of-the-art results in measuring semantic distance. These results also support the broader claim that natural language problems in a resource-poor language can be solved using a knowledge source in a resource-rich language (for example the cross-lingual PoS tagger of Cucerzan and Yarowsky (2002)). Cross-lingual DPCs also have tremendous potential in tasks inherently involving more than one language. We believe that the future of natural language processing lies not in standalone monolingual systems but in those that are powered by automatically created multilingual networks of information.

Thus distributional measures of concept-distance have most of the attractive features of a distributional measure, and yet avoid to some extent the problems associated with word-distance measures. They do not conflate sense information (mitigating the limitation described in Section 3.2.1). They need a matrix of size only about 1000 by 1000 to store all pre-computed distances (mitigating limitation of Section 3.3). As they calculate distance between coarse senses, each represented by many words, even if some words are not seen often in a text corpus, all concepts have sufficient representation even in small corpora, they avoid the data sparseness problem (limitation of Section 3.2.2). And lastly, they can be used cross-lingually so that a high-quality knowledge source in a resource-rich language can be leveraged to solve semantic distance problems in a resource-poor language (mitigating the limitation described in 3.1.1).

### Acknowledgments

This paper incorporates research that was first reported in Mohammad and Hirst (2006a), Mohammad and Hirst (2006b), and Mohammad et al. (2007). The work described in section 7 was carried out in collaboration with Iryna Gurevych and Torsten Zesch, Technische Universität Darmstadt. The research was supported by the Natural Sciences and Engineering Research Council of Canada, the University of Toronto, the U.S. National Science Foundation, and 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. We thank Afra Alishahi, Alex Budanitsky, Michael Demko, Afsaneh Fazly, Diana McCarthy, Rada Mihalcea, Siddharth Patwardhan, Gerald Penn, Philip Resnik, Frank Rudicz, Suzanne Stevenson, Vivian Tsang, and Xinglong Wang for helpful discussions.

### References

- Agirre, Eneko 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.
- Bernard, John R. L., editor. 1986. *The Macquarie Thesaurus*. Macquarie Library, Sydney, Australia.
- Budanitsky, Alexander and Graeme Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2000)*, Pittsburgh, Pennsylvania.
- Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.
- Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.
- Cruse, D. Allen. 1986. *Lexical semantics*. Cambridge University Press, Cambridge, UK.
- Cucerzan, Silviu and David Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of the 6th Conference on Computational Natural Language Learning*, pages 132–138, Taipei, Taiwan.
- Curran, James R. 2004. *From Distributional to Semantic Similarity*. Ph.D. thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.
- Dagan, Ido, Lillian Lee, and Fernando Pereira. 1994. Similarity-based estimation of word cooccurrence probabilities. In *Proceedings of the 32nd Annual Meeting of the Association of Computational Linguistics (ACL-1994)*, pages 272–278, Las Cruces, New Mexico.
- Erk, Katrin 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.
- Finkelstein, Lev, Evgeniy Gabrilovich, 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.

- Firth, John R. 1957. A synopsis of linguistic theory 1930–55. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32, Oxford, England. The Philological Society.
- Gurevych, Iryna. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-2005)*, pages 767–778, Jeju Island, Republic of Korea.
- Harris, Zellig. 1968. *Mathematical Structures of Language*. Interscience Publishers, New York, NY.
- Hirst, Graeme and Alexander Budanitsky. 2005. Correcting real-word spelling errors by restoring lexical cohesion. *Natural Language Engineering*, 11(1):87–111.
- Hirst, Graeme and David St-Onge. 1998. Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, chapter 13, pages 305–332.
- Jarmasz, Mario and Stan Szpakowicz. 2003. Roget'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.
- Jiang, Jay J. and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research on Computational Linguistics (ROCLING X)*, Taipei, Taiwan.
- Landauer, Thomas K. and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.
- Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25(2–3):259–284.
- Leacock, Claudia and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, MA, chapter 11, pages 265–283.
- Lee, Lillian. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics (AISTATS-2001)*, pages 65–72, Key West, Florida.
- Lin, Dekang. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL, EACL-1997)*, pages 64–71, Madrid, Spain.
- Lin, Dekang. 1998a. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998)*, pages 768–773, Montreal, Canada.
- Lin, Dekang. 1998b. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304, San Francisco, CA. Morgan Kaufmann.
- Manning, Christopher D. and Hinrich Schütze. 2008. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.
- Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.
- Mohammad, Saif, 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.
- Mohammad, Saif and Graeme Hirst. 2006a. Determining word sense dominance using a thesaurus. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 121–128, Trento, Italy.
- Mohammad, Saif and Graeme Hirst. 2006b. Distributional measures of concept-distance: A task-oriented evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pages 35–43, Sydney, Australia.
- Mohammad, Saif and Graeme Hirst. 2007. Distributional measures of semantic distance: A survey. <http://www.cs.toronto.edu/compling/Publications>.
- Mohammad, Saif, Graeme Hirst, and Philip Resnik. 2007. Tor, tormd: Distributional profiles of concepts for unsupervised word sense disambiguation. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-07)*, pages 326–333, Prague, Czech Republic.
- Morris, Jane 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.
- Navigli, Roberto. 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 for Computational Linguistics (COLING-ACL 2006)*, pages 105–112, Sydney, Australia.
- Pantel, Patrick. 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.
- Pantel, Patrick and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of the 8th Association of Computing Machinery SIGKDD International Conference On Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.
- Patwardhan, Siddharth, Satanjeev Banerjee, and Ted Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03)*, pages 17–21, Mexico City, Mexico.
- Patwardhan, Siddharth and Ted Pedersen. 2006. Using WordNet based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the European Chapter of the Association for Computational Linguistics Workshop Making Sense of Sense – Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8, Trento, Italy.
- Rada, Roy, Hafedh Mili, Ellen Bicknell, and Maria Blettner. 1989. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30.
- Resnik, Philip. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pages 448–453, Montreal, Canada.
- Resnik, Philip. 1998. Wordnet and class-based probabilities. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts, pages 239–263.
- Resnik, Philip. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Communications of the Association of Computing Machinery*, 11:95–130.
- Resnik, Philip and Mona Diab. 2000. Measuring verb similarity. In *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society (CogSci 2000)*, pages 399–404, Philadelphia, Pennsylvania.
- Rubenstein, Herbert and John B. Goodenough. 1965a. Contextual correlates of synonymy. *Communications of the Association of Computing Machinery*, 8(10):627–633.
- Rubenstein, Herbert and John B. Goodenough. 1965b. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633.
- Schütze, Hinrich and Jan O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing and Management*, 33(3):307–318.
- Turney, Peter. 2001. Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, pages 491–502, Freiburg, Germany.
- Turney, Peter. 2006. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 313–320, Sydney, Australia.
- Véronis, Jean. 2004. Hyperlex: Lexical cartography for information retrieval. *Computer Speech and Language. Special Issue on Word Sense Disambiguation*, 18(3):223–252.
- Wallace, DeWitt and Lila Acheson Wallace. 2005. *Reader's Digest, das Beste für Deutschland*. Jan 2001–Dec 2005. Verlag Das Beste, Stuttgart.
- Weeds, Julie, 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.
- Yarowsky, David. 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-1992)*, pages 454–460, Nantes, France.
- Zesch, Torsten, Iryna Gurevych, and Max Mühlhäuser. 2007. Comparing Wikipedia and German WordNet by evaluating semantic relatedness on multiple datasets. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL,HLT-2007)*, pages 205–208, Rochester, New York.

|

|

—

—

—

—