Cross-lingual Distributional Profiles of Concepts for Measuring Semantic Distance

Saif Mohammad[†] **Iryna Gurevych**[*] **Graeme Hirst**[†] **Torsten Zesch**[*]

[†]Dept. of Computer Science
University of Toronto
Toronto, Canada
{smm,gh}@cs.toronto.edu

[*]Ubiquitous Knowledge Processing Group
Darmstadt University of Technology
Darmstadt, Germany
{gurevych,zesch}@tk.informatik.tu-darmstadt.de

Abstract

We present the idea of estimating semantic distance in one, possibly resource-poor, language using a knowledge source in another, possibly resource-rich, language. We do so by creating cross-lingual distributional profiles of concepts, using a bilingual lexicon and a bootstrapping algorithm, but without the use of any sense-annotated data or word-aligned corpora. The cross-lingual measures of semantic distance are evaluated on two tasks: (1) estimating semantic distance between words and ranking the word pairs according to semantic distance, and (2) solving *Reader's Digest* 'Word Power' problems. In task (1), cross-lingual measures are superior to conventional monolingual measures based on a wordnet. In task (2), cross-lingual measures are able to solve more problems correctly, and despite scores being affected by many tied answers, their overall performance is again better than the best monolingual measures.

1 Introduction

Accurately estimating the semantic distance between concepts or between words in context has pervasive applications in computational linguistics, including machine translation, information retrieval, speech recognition, spelling correction, and text categorization (see Budanitsky and Hirst (2006) for discussion), and it is becoming clear that basing such measures on a combination of corpus statistics with a knowledge source, such as a dictionary, published thesaurus, or WordNet, can result in higher accuracies (Mohammad and Hirst, 2006b). This is because such knowledge sources capture semantic information about concepts and, to some extent, world knowledge. They also act as sense inventories for the words in a language.

However, applying algorithms for semantic distance to most languages is hindered by the lack of linguistic resources. In this paper, we propose a new method that allows us to compute semantic distance in a possibly resource-poor language by seamlessly combining its text with a knowledge source in a different, preferably resource-rich, language. We demonstrate the approach by combining German text with an English thesaurus to create English–German distributional profiles of concepts, which in turn will be used to measure the semantic distance between German words.

Two classes of methods have been used in determining semantic distance. **Semantic measures of concept-distance**, such as those of Jiang and Conrath (1997) and Resnik (1995), rely on the structure of a knowledge source, such as WordNet, to determine the distance between two concepts defined in it (see Budanitsky and Hirst (2006) for a survey). **Distributional measures of word-distance**[1], such as cosine and α-skew divergence (Lee, 2001), deem

[1] Many distributional approaches represent the sets of contexts of the target words as points in multidimensional co-occurrence space or as co-occurrence distributions. A measure, such as cosine, that captures vector distance or a measure, such as α-skew divergence, that captures distance between distributions is then used to measure distributional distance. We will therefore refer to these measures as distributional measures.

two words to be closer or less distant if they occur in similar contexts (see Mohammad and Hirst (2005) for a comprehensive survey).

Distributional measures rely simply on raw text and possibly some shallow syntactic processing. They do not require any other manually-created resource, and tend to have a higher coverage. However, by themselves they perform poorly when compared to semantic measures (Mohammad and Hirst, 2006b) because when given a target word pair we usually need the distance between their closest senses, but distributional measures of word-distance tend to conflate the distances between all possible sense pairs. Latent semantic analysis (LSA) (Landauer et al., 1998) has also been used to measure distributional distance with encouraging results (Rapp, 2003). However, it too measures the distance between words and not senses. Further, the dimensionality reduction inherent to LSA has the effect of making the predominant sense more dominant while de-emphasizing the other senses. Therefore, an LSA-based approach will also conflate information from the different senses, and even more emphasis will be placed on the predominant senses. Given the semantically close target nouns *play* and *actor*, for example, a distributional measure will give a score that is some sort of a dominance-based average of the distances between their senses. The noun *play* has the predominant sense of ‘children’s recreation’ (and not ‘drama’), so a distributional measure will tend to give the target pair a large (and thus erroneous) distance score. Also, distributional word-distance approaches need to create large $V \times V$ co-occurrence and distance matrices, where V is the size of the vocabulary (usually at least 100,000).[2]

Mohammad and Hirst (2006b) proposed a way of combining written text with a published thesaurus to measure distance between *concepts* (or word senses) using distributional measures, thereby eliminating sense-conflation and achieving results better than the simple word-distance measures and indeed also most of the WordNet-based semantic measures. We called these measures **distributional measures of concept-distance**. Concept-distance measures can be used to measure distance between a word pair by choosing the distance between their closest senses. Thus, even though ‘children’s recreation’ is the predominant sense of *play*, the ‘drama’ sense is much closer to *actor* and so their distance will be chosen. These distributional concept-distance approaches need to create only $V \times C$ co-occurrence and $C \times C$ distance matrices, where C is the number of categories or senses (usually about 1000). It should also be noted that unlike the best WordNet-based measures, distributional measures (both word- and concept-distance ones) can be used to estimate not just semantic similarity but also semantic relatedness—useful in many tasks including information retrieval. However, the high-quality thesauri and (to a much greater extent) WordNet-like resources that these methods require do not exist for most of the 3000–6000 languages in existence today and they are costly to create.

In this paper, we introduce **cross-lingual distributional measures of concept-distance**, or simply **cross-lingual measures**, that determine the distance between a word pair belonging to a resource-poor language using a knowledge source in a resource-rich language and a bilingual lexicon[3]. We will use the cross-lingual measures to calculate distances between German words using an English thesaurus and a German corpus. Although German is not resource-poor *per se*, Gurevych (2005) has observed that the German wordnet GermaNet (Kunze, 2004) (about 60,000 synsets) is less developed than the English WordNet (Fellbaum, 1998) (about 117,000 synsets) with respect to the coverage of lexical items and lexical semantic relations represented therein. On the other hand, substantial raw corpora are available for the German language. Crucially for our evaluation, the existence of GermaNet allows comparison of our cross-lingual approach with monolingual ones.

2 Monolingual Distributional Measures

In order to set the context for cross-lingual concept-distance measures (Section 3), we first summarize monolingual distributional approaches, with a focus on distributional concept-distance measures.

[2] LSA is especially expensive as singular value decomposition, a key component for dimensionality reduction, requires computationally intensive matrix operations; making it less scalable to large amounts of text (Gorman and Curran, 2006).

[3] For most languages that have been the subject of academic study, there exists at least a bilingual lexicon mapping the core vocabulary of that language to a major world language and a corpus of at least a modest size.

2.1 Word-distance

Words that occur in similar contexts tend to be semantically close. In our experiments, we defined the context of a target word, its co-occurring words, to be ±5 words on either side (but not crossing sentence boundaries). The set of contexts of a target word is usually represented by the strengths of association of the target with its co-occurring words, which we refer to as the **distributional profile (DP)** of the word. Here is a constructed example DP of the word *star*:

> **DP of a word**
> ***star***: *space* 0.28, *movie* 0.2, *famous* 0.13, *light* 0.09, *rich* 0.04, . . .

Simple counts are made of how often the target word co-occurs with other words in text and how often the words occur individually. A suitable statistic, such as pointwise mutual information (PMI), is then applied to these counts to determine the strengths of association between the target and co-occurring words. The distributional profiles of two target words represent their contexts as points in multi-dimensional word-space. A suitable distributional measure (for example, cosine) gives the distance between the two points, and thereby an estimate of the semantic distance between the target words.

2.2 Concept-distance

In Mohammad and Hirst (2006b), we show how distributional profiles of *concepts* (DPCs) can be used to measure semantic distance. Below are the DPCs or DPs of two senses of the word *star* (the senses or concepts themselves are glossed by a set of near-synonymous words, placed in parentheses):

> **DPs of concepts**
> **‘celestial body’** (*celestial body, sun, . . .*): *space* 0.36, *light* 0.27, *constellation* 0.11, . . .
> **‘celebrity’** (*celebrity, hero, . . .*): *famous* 0.24, *movie* 0.14, *rich* 0.14, . . .

Thus the profiles of two target *concepts* represent their contexts as points in multi-dimensional word-space. A suitable distributional measure (for example, cosine) can then be used to give the distributional distance between the two concepts in the same way that distributional word-distance is measured.

But to calculate the strength of association of a concept with co-occurring words, in order to create DPCs, we must determine the number of times a word used in that sense co-occurs with surrounding words. In Mohammad and Hirst (2006a), we proposed a way to determine these counts without the use of sense-annotated data. Briefly, a **word–category co-occurrence matrix (WCCM)** is created having English word types w^{en} as one dimension and English thesaurus categories c^{en} as another. We used the *Macquarie Thesaurus* (Bernard, 1986) both as a very coarse-grained sense inventory and a source of possibly ambiguous English words that together unambiguously represent each category (concept). The WCCM is populated with co-occurrence counts from a large English corpus (we used the *British National Corpus (BNC)*). A particular cell m_{ij}, corresponding to word w_i^{en} and concept c_j^{en}, is populated with the number of times w_i^{en} co-occurs (in a window of ±5 words) with any word that has c_j^{en} as one of its senses (i.e., w_i^{en} co-occurs with any word listed under concept c_j^{en} in the thesaurus).

$$\begin{array}{c|ccccc} & c_1^{en} & c_2^{en} & \dots & c_j^{en} & \dots \\ \hline w_1^{en} & m_{11} & m_{12} & \dots & m_{1j} & \dots \\ w_2^{en} & m_{21} & m_{22} & \dots & m_{2j} & \dots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ w_i^{en} & m_{i1} & m_{i2} & \dots & m_{ij} & \dots \\ \vdots & \vdots & \vdots & \dots & \vdots & \ddots \end{array}$$

This matrix, created after a first pass of the corpus, is the **base word–category co-occurrence matrix (base WCCM)** and it captures strong associations between a sense and co-occurring words.[4] This is similar to how Yarowsky (1992) identifies words that are indicative of a particular sense of the target.

We know that words that occur close to a target word tend to be good indicators of its intended sense. Therefore, we make a second pass of the corpus, using the base WCCM to roughly disambiguate the words in it. For each word, the strength of association of each of the words in its context (±5 words)

[4] From the base WCCM we can determine the number of times a word w and concept c co-occur, the number of times w co-occurs with any concept, and the number of times c co-occurs with any word. A statistic such as PMI can then give the strength of association between w and c.

with each of its senses is summed. The sense that has the highest cumulative association is chosen as the intended sense. A new **bootstrapped WCCM** is created such that each cell m_{ij}, corresponding to word w_i^{en} and concept c_j^{en}, is populated with the number of times w_i^{en} co-occurs with any word *used in sense* c_j^{en}.

Mohammad and Hirst (2006a) used the DPCs created from the bootstrapped WCCM to attain near-upper-bound results in the task of determining word sense dominance. Unlike the McCarthy et al. (2004) dominance system, our approach can be applied to much smaller target texts (a few hundred sentences) without the need for a large similarly-sense-distributed text[5]. In Mohammad and Hirst (2006a), the DPC-based monolingual distributional measures of *concept-distance* were used to rank word pairs by their semantic similarity and to correct real-word spelling errors, attaining markedly better results than monolingual distributional measures of *word-distance*. In the spelling correction task, the distributional concept-distance measures performed better than all WordNet-based measures as well, except for the Jiang and Conrath (1997) measure.

3 Cross-lingual Distributional Measures

We now describe how distributional measures of concept-distance can be used in a cross-lingual framework to determine the distance between words in (resource-poor) language L_1 by combining its text with a thesaurus in (resource-rich) language L_2, using an L_1–L_2 bilingual lexicon. We will compare this approach with the best monolingual approaches; the smaller the loss in performance, the more capable the 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 monolingual evaluation in German will use GermaNet. The remainder of the paper describes our approach in terms of German and English, but the algorithm itself is language independent.

3.1 Concept-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 **cross-lingual candidate senses** of the German word w^{de}.[6] Figure 1 depicts examples.[7]

As in the monolingual distributional measures, the distance between two concepts is calculated by first determining their DPs. However, 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. Here are constructed example cross-lingual DPs of the two senses of *star*:

> **Cross-lingual DPs of concepts**
> **'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, . . .

In order to calculate the strength of association, we must first determine individual word and concept counts, as well as their co-occurrence counts.

3.2 Cross-lingual word–category co-occurrence matrix

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}

[5] The McCarthy et al. (2004) system needs to first generate a distributional thesaurus from the target text (if it is large enough—a few million words) or from another large text with a distribution of senses similar to the target text.

[6] Some of the cross-lingual candidate senses of w^{de} might not really be senses of w^{de} (e.g., 'celebrity', 'river bank', and 'judiciary' in Figure 1). However, as substantiated by experiments in Section 4, our algorithm is able to handle the added ambiguity.

[7] Vocabulary of German words needed to understand this discussion: ***Bank***: 1. financial institution, 2. bench (furniture); ***berühmt***: famous; ***Film***: movie (motion picture); ***Himmelskörper***: heavenly body; ***Konstellation***: constellation; ***Licht***: light; ***Morgensonne***: morning sun; ***Raum***: space; ***reich***: rich; ***Sonne***: sun; ***Star***: star (celebrity); ***Stern***: star (celestial body)

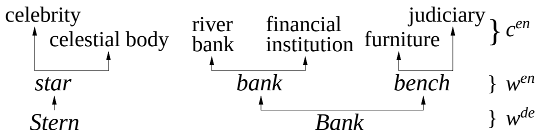

Figure 1: The cross-lingual candidate senses of German words *Stern* and *Bank*.

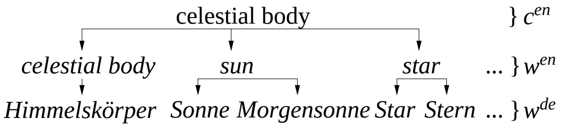

Figure 2: Words having ‘celestial body’ as one of their cross-lingual candidate senses.

as another.

$$\begin{array}{c|ccccc} & c_1^{en} & c_2^{en} & \cdots & c_j^{en} & \cdots \\ \hline w_1^{de} & m_{11} & m_{12} & \cdots & m_{1j} & \cdots \\ w_2^{de} & m_{21} & m_{22} & \cdots & m_{2j} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ w_i^{de} & m_{i1} & m_{i2} & \cdots & m_{ij} & \cdots \\ \vdots & \vdots & \vdots & \cdots & \vdots & \ddots \end{array}$$

The matrix is populated with co-occurrence counts from a large German corpus; we used the newspaper corpus, *taz*[8] (Sep 1986 to May 1999; 240 million words). 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 ±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 2). We used the *Macquarie Thesaurus* (Bernard, 1986) (about 98,000 words) for our experiments. The possible German translations of an English word were taken from the German–English bilingual lexicon BEOLINGUS[9] (about 265,000 entries).

This base word–category co-occurrence matrix (base WCCM), created after a first pass of the corpus captures strong associations between a category (concept) and co-occurring words. 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 Figure 2), 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’ will be relatively higher than that of *Raum* and ‘celebrity’.

As in the monolingual case, a second pass of the corpus is made to disambiguate the (German) words in it. For each word, the strength of association of each of the words in its context (±5 words) with each of its cross-lingual candidate senses is summed. The sense that has the highest cumulative association with co-occurring words is chosen as the intended sense. A new bootstrapped WCCM is created by populating each cell m_{ij}, corresponding to word w_i^{de} and concept c_j^{en}, with the number of times the German word w_i^{de} co-occurs with any German word *used in cross-lingual sense* c_j^{en}. A statistic such as PMI is then applied to these counts to determine the strengths of association between a target concept and co-occurring words, giving the distributional profile of the concept.

Following the ideas described above, Mohammad et al. (2007) created Chinese–English DPCs from Chinese text, a Chinese–English bilingual lexicon, and an English thesaurus. They used these DPCs to implement an unsupervised naïve Bayes word sense classifier that placed first among all unsupervised systems taking part in the Multilingual Chinese–English Lexical Sample Task (task #5) of SemEval-07 (Jin et al., 2007).

4 Evaluation

We evaluated the newly proposed 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 ‘Word Power’ questions from *Reader’s Digest*. In order to compare results with state-of-the-art monolingual approaches we conducted experiments using Ger-

[8] http://www.taz.de

[9] http://dict.tu-chemnitz.de

(Cross-lingual) Distributional Measures	**(Monolingual) GermaNet Measures**	
	Information Content–based	**Lesk-like**
α-skew divergence (Lee, 2001) (***ASD***)	Jiang and Conrath (1997) (***JC***)	hypernym pseudo-gloss (***HPG***)
cosine (Schütze and Pedersen, 1997) (***Cos***)	Lin (1998b) (***Lin$_{GN}$***)	radial pseudo-gloss (***RPG***)
Jensen-Shannon divergence (***JSD***)	Resnik (1995) (***Res***)	
Lin's measure (1998a) (***Lin$_{dist}$***)		

Table 1: Distance measures used in our experiments.

Dataset	**Year**	**Language**	**# pairs**	**PoS**	**Scores**	**# subjects**	**Correlation**
Gur65	2005	German	65	N	discrete {0,1,2,3,4}	24	.810
Gur350	2006	German	350	N, V, A	discrete {0,1,2,3,4}	8	.690

Table 2: Comparison of datasets used for evaluating semantic distance in German.

maNet measures as well. The specific distributional measures[10] and GermaNet-based measures we used are listed in Table 1. The GermaNet measures are of two kinds: (1) information content measures,[11] and (2) Lesk-like measures that rely on *n*-gram overlaps in the glosses of the target senses, proposed by Gurevych (2005)[12].

The cross-lingual measures combined the German newspaper corpus *taz* with the English *Macquarie Thesaurus* using the German–English bilingual lexicon BEOLINGUS. Multi-word expressions in the thesaurus and the bilingual lexicon were ignored. We used a context of ±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.

4.1 Measuring distance in word pairs

4.1.1 Data

A direct approach to evaluate distance measures is to compare them with human judgments. 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**) consists of a German translation of the English Rubenstein and Goodenough (1965) 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**[13] are summarized in Table 2. Inter-subject correlations are indicative of the degree of ease in annotating the datasets.

4.1.2 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 ρ; 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 be-

[10] JSD and ASD calculate the difference in distributions of words that co-occur with the targets. *Lin$_{dist}$* (distributional measure) and *Lin$_{GN}$* (GermaNet measure) follow from Lin's (1998b) information-theoretic definition of similarity.

[11] 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. In general, 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.

[12] 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.

[13] The datasets are publicly available at: http://www.ukp.tu-darmstadt.de/data/semRelDatasets

tween 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 the Pearson correlations with caution.

Table 3 shows the results.[14] Observe that on both datasets and by both measures of correlation, cross-lingual measures of concept-distance perform not just as well as the best monolingual measures, but in fact better. 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 2). Considering Spearman's rank correlation, α-skew divergence and Jensen-Shannon divergence perform best on both datasets. The correlations of cosine and Lin_{dist} are not far behind. Amongst the monolingual GermaNet measures, radial pseudo-gloss performs best. Considering Pearson's correlation, Lin_{dist} performs best overall and radial pseudo-gloss does best amongst the monolingual measures. Thus, we see that on both datasets and as per both measures of correlation, the cross-lingual measures perform not just as well as the best monolingual measures, but indeed slightly better.

4.2 Solving word choice problems from *Reader's Digest*

4.2.1 Data

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. The correct answer may be a near-synonym of the target or it may be related to the target by some other classical or non-classical relation (usually the former). For example:[15]

Duplikat (duplicate)

a. *Einzelstück* (single copy)	b. *Doppelkinn* (double chin)
c. *Nachbildung* (replica)	d. *Zweitschrift* (copy)

Our 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.

The **Reader's Digest Word Power (RDWP) benchmark** for German 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). We discarded 44 problems that had more than one correct answer, and 20 problems that used a phrase instead of a single term as the target. 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. We assign 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[16] as the target, then only these are chosen as the closest senses.

The German RDWP dataset contains many phrases that cannot be found in the knowledge sources (GermaNet or *Macquarie Thesaurus* via translation list). In these cases, we remove stop-

[14] In Table 3, all values are statistically significant at the 0.01 level (2-tailed), except for the one in italic (*0.212*), which is significant at the 0.05 level (2-tailed).

[15] English translations are in parentheses.

[16] 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.

	Gur65		**Gur350**	
Measure	ρ	*r*	ρ	*r*
Monolingual				
HPG	0.672	0.702	0.346	0.331
RPG	**0.764**	0.565	**0.492**	0.420
JC	0.665	**0.748**	0.417	0.410
Lin_{GN}	0.607	0.739	0.475	**0.495**
Res	0.623	0.722	0.454	0.466
Cross-lingual				
ASD	**0.794**	0.597	**0.520**	0.413
Cos	0.778	0.569	0.500	*0.212*
JSD	**0.793**	0.633	**0.522**	0.422
Lin_{dist}	0.775	**0.816**	0.498	**0.514**

Table 3: Correlations of distance measures with human judgments.

	Reader's Digest Word Power benchmark						
Measure	**Att.**	**Cor.**	**Ties**	**Score**	**P**	**R**	**F**
Monolingual							
HPG	222	174	11	**171.5**	.77	.17	**.28**
RPG	266	188	15	**184.7**	.69	.18	**.29**
JC	357	157	1	156.0	.44	.16	.23
Lin_{GN}	298	153	1	152.5	.51	.15	.23
Res	299	154	33	148.3	.50	.15	.23
Cross-lingual							
ASD	438	185	81	151.6	.35	.15	.21
Cos	438	276	90	**223.1**	.51	.22	**.31**
JSD	438	276	90	**229.6**	.52	.23	**.32**
Lin_{dist}	438	274	90	**228.7**	.52	.23	**.32**

Table 4: Performance of distance measures on word choice problems. (Att.: Attempted, Cor.: Correct)

words (prepositions, articles, etc.) and split the phrase into component words. As German words in a phrase can be highly inflected, we lemmatize all components. 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*. We compute semantic distance between the target and each phrasal component and select the minimum value as the distance between target and potential answer.

4.2.2 Results and Discussion

Table 4 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 'Att.' 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 answers too (column 'Cor.'), but this number is bloated due to the large number of ties.[17] 'Score' is the score each measure gets after it is penalized for the ties. The cross-lingual measures *Cos*, *JSD*, and Lin_{dist} 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, recall, and F-scores ($P = Score/Att$; $R = Score/1008$; $F = 2 \times P \times R/(P+R)$). Observe that the cross-lingual measures have a higher coverage (recall) than the monolingual measures but lower precision. The F scores show that the best cross-lingual measures do slightly better than the best monolingual ones, despite the large number of ties. The measures of *Cos*, *JSD*, and Lin_{dist} remain the best cross-lingual measures, whereas HPG and RPG are the best monolingual ones.

5 Conclusion

We have proposed a new method to determine semantic distance in a possibly resource-poor language by combining its text with a knowledge source in a different, preferably resource-rich, language. Specifically, we combined German text with an English thesaurus to create cross-lingual distributional profiles of concepts—the strengths of association between English thesaurus senses (concepts) of German words and co-occurring German words—using a German–English bilingual lexicon and a bootstrapping algorithm designed to overcome ambiguities of word-senses and translations. Notably, we do so without the use of sense-annotated text or word-aligned parallel corpora. We did not parse or chunk the text, nor did we stem, lemmatize, or part-of-speech-tag the words.

We used the cross-lingual DPCs to estimate semantic distance by developing new cross-lingual

[17] 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 cross-lingual measures operate on the fine-grained GermaNet.

distributional measures of concept-distance. These measures are like the distributional measures of concept-distance (Mohammad and Hirst, 2006a, 2006b), except they can determine distance between words in one language using a thesaurus in a different language. 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 German semantic distance benchmarks. In the process, we compiled a large German benchmark of *Reader's Digest* word choice problems suitable for evaluating semantic-relatedness measures. Most previous semantic distance benchmarks are either much smaller or cater primarily to semantic similarity measures.

Even with the added ambiguity of translating words from one language to another, the cross-lingual measures performed 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 achieved 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 achieving 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 our 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 and far-reaching claim that natural language problems in a resource-poor language can be solved using a knowledge source in a resource-rich language (e.g., Cucerzan and Yarowsky's (2002) cross-lingual PoS tagger). Our future work will explore other tasks such as information retrieval and text categorization. Cross-lingual DPCs also have tremendous potential in tasks inherently involving more than one language, such as machine translation and multi-language multi-document summarization. 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.

Acknowledgments

We thank Philip Resnik, Michael Demko, Suzanne Stevenson, Frank Rudicz, Afsaneh Fazly, and Afra Alishahi for helpful discussions. This research is financially supported by the Natural Sciences and Engineering Research Council of Canada, the University of Toronto, the German Research Foundation under the grant "Semantic Information Retrieval" (SIR), GU 798/1-2.

References

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

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Silviu Cucerzan 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.

Christiane Fellbaum. 1998. *WordNet An Electronic Lexical Database*. MIT Press, Cambridge, MA.

James Gorman and James R. Curran. 2006. Scaling distributional similarity to large corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 361–368, Sydney, Australia.

Iryna Gurevych. 2005. Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pages 767–778, Jeju Island, Republic of Korea.

Mario Jarmasz 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.

Jay J. Jiang 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)*, Taiwan.

Peng Jin, Yunfang Wu, and Shiwen Yu. 2007. SemEval-2007 task 05: Multilingual Chinese-English lexical sample task. In *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SemEval-07)*, Prague, Czech Republic.

Claudia Kunze, 2004. *Lexikalisch-semantische Wortnetze*, chapter Computerlinguistik und Sprachtechnologie, pages 423–431. Spektrum Akademischer Verlag.

Thomas K. Landauer 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.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25(2–3):259–284.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pages 65–72.

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

Dekang Lin. 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.

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

Saif Mohammad and Graeme Hirst. 2005. Distributional measures as proxies for semantic relatedness. *In submission*, http://www.cs.toronto.edu/compling/Publications.

Saif Mohammad 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)*, Trento, Italy.

Saif Mohammad 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)*, Sydney, Australia.

Saif Mohammad, Graeme Hirst, and Philip Resnik. 2007. Distributional profiles of concepts for unsupervised word sense disambigution. In *Proceedings of the Fourth International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SemEval-07)*, Prague, Czech Republic.

Jane Morris and Graeme Hirst. 2004. Non-classical lexical semantic relations. In *Proceedings of the Workshop on Computational Lexical Semantics, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Massachusetts.

Reinhard Rapp. 2003. Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the Machine Translation Summit IX*, pages 315–322, New Orleans, Louisiana.

Philip Resnik. 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.

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

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

Peter Turney. 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.

Peter Turney. 2006. Expressing implicit semantic relations without supervision. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 313–320, Sydney, Australia.

DeWitt Wallace and Lila Acheson Wallace. 2005. *Reader's Digest, das Beste für Deutschland*. Jan 2001–Dec 2005. Verlag Das Beste, Stuttgart.

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

Torsten Zesch, 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.