A General Framework for Distributional Similarity

Julie Weeds and David Weir
School of Cognitive and Computing Sciences
University of Sussex
Brighton, BN1 9QH, UK
{juliewe, davidw}@cogs.susx.ac.uk

Abstract

We present a general framework for distributional similarity based on the concepts of precision and recall. Different parameter settings within this framework approximate different existing similarity measures as well as many more which have, until now, been unexplored. We show that optimal parameter settings outperform two existing state-of-the-art similarity measures on two evaluation tasks for high and low frequency nouns.

1 Introduction

There are many potential applications of sets of distributionally similar words. In the syntactic domain, language models, which can be used to evaluate alternative interpretations of text and speech, require probabilistic information about words and their co-occurrences which is often not available due to the sparse data problem. In order to overcome this problem, researchers (e.g. Pereira et al. (1993)) have proposed estimating probabilities based on sets of words which are known to be distributionally similar. In the semantic domain, the hypothesis that words which mean similar things behave in similar ways (Levin, 1993), has led researchers (e.g. Lin (1998)) to propose that distributional similarity might be used as a predictor of semantic similarity. Accordingly, we might automatically build thesauruses which could be used in tasks such as malapropism correction (Budanitsky and Hirst, 2001) and text summarization (Silber and McCoy, 2002).

However, the loose definition of distributional similarity — that two words are distributionally similar if they appear in similar contexts — has led to many distributional similarity measures being proposed; for example, the L_1 Norm, the Euclidean Distance, the Cosine Metric (Salton and McGill, 1983), Jaccard's Coefficient (Frakes and Baeza-Yates, 1992), the Dice Coefficient (Frakes and Baeza-Yates, 1992), the Kullback-Leibler Divergence (Cover and Thomas, 1991), the Jenson-Shannon Divergence (Rao, 1983), the α-skew Divergence (Lee, 1999), the Confusion Probability (Essen and Steinbiss, 1992), Hindle's Mutual Information(MI)-Based Measure (Hindle, 1990) and Lin's MI-Based Measure (Lin, 1998).

Further, there is no clear way of deciding which is the best measure. Application-based evaluation tasks have been proposed, yet it is not clear (Weeds and Weir, 2003) whether there is or should be one distributional similarity measure which outperforms all other distributional similarity measures on all tasks and for all words.

We take a generic approach that does not directly reduce distributional similarity to a single dimension. The way dimensions are combined together will depend on parameters tuned to the demands of a given application. Further, different parameter settings will approximate different existing similarity measures as well as many more which have, until now, been unexplored.

The contributions of this paper are four-fold. First, we propose a general framework for distributional similarity based on the concepts of precision and recall (Section 2). Second, we evaluate the framework at its optimal parameter settings for two different applications (Section 3), showing that it outperforms existing state-of-the-art similarity measures for both high and low frequency nouns. Third, we begin to investigate to what extent existing similarity measures might be characterised in terms of parameter settings within the framework (Section 4). Fourth, we provide an understanding of why a single existing measure cannot achieve optimal results in every application of distributional similarity measures.

2 The Framework

In this section, we introduce the relevance of the Information Retrieval (IR) concepts of precision and recall in the context of word similarity. We provide combinatorial, probabilistic and mutual-information based models for precision and recall and discuss combining precision and recall to provide a single number in the context of a particular application.

2.1 Precision and Recall

The similarity[1] of two nouns can be viewed as a measure of how appropriate it is to use one noun (or its distribution) in place of the other. If we are using the distribution of one noun in place of the distribution the other noun, we can consider the precision and recall of the prediction made. Precision tells us how much of what has been predicted is correct whilst recall tells us how much of what is required has been predicted.

In order to calculate precision and recall, we first need to consider for each noun n which verb co-occurrences will be predicted by it and, conversely, required in a description of it. We will refer to these verbs as the features of n, $F(n)$:

$$F(n) = \{v : D(n,v) > 0\}$$

where $D(n,v)$, is the degree of association between noun n and verb v. Possible association functions will be defined in the context of each model described below.

If we are considering the ability of noun A to predict noun B then it follows that the set of True Positives is $TP = F(A) \cap F(B)$ and precision and recall can be defined as:

$$\begin{aligned}\mathcal{P}(A,B) &= \frac{\sum_{TP} D(A,v)}{\sum_{F(A)} D(A,v)} \\ \mathcal{R}(A,B) &= \frac{\sum_{TP} D(B,v)}{\sum_{F(B)} D(B,v)}\end{aligned}$$

Precision and recall both lie in the range [0,1] and are both equal to one when each noun has exactly the same features. It should also be noted that $\mathcal{R}(A,B) = \mathcal{P}(B,A)$.

We will now consider some different possibilities for measuring the degree of association between a noun n and a verb v.

2.2 Combinatorial Model

In the combinatorial model, we simply consider whether a verb has ever been seen to co-occur with the noun. In other words, the degree of association (D) between a noun n and a verb v is 1 if they have co-occurred together and 0 otherwise.

$$D_c(n,v) = \begin{cases} 1 \text{ if } P(v|n) > 0 \\ 0 \text{ otherwise} \end{cases}$$

In this case, it should be noted that the definitions of precision and recall can be simplified as follows:

$$\begin{aligned}\mathcal{P}_c(A,B) &= \frac{\sum_{TP} D_c(A,v)}{\sum_{F(A)} D_c(A,v)} = \frac{|TP|}{|F(A)|} \\ \mathcal{R}_c(A,B) &= \frac{\sum_{TP} D_c(B,v)}{\sum_{F(B)} D_c(B,v)} = \frac{|TP|}{|F(B)|}\end{aligned}$$

2.3 Probabilistic Model

In the probabilistic model, more probable (or more frequent) co-occurrences are considered more significant. The degree of association between a noun n and verb v is defined in the probabilistic model as:

$$D_p(n,v) = P(v|n)$$

The definitions for feature set membership, TP, precision and recall all remain the same except for the use of the new association function.

Using the probabilistic model, the precision of A's prediction of B is the probability that a verb picked at random from those co-occurring with A will also co-occur with B; and the recall of A's prediction of B is the probability that a verb picked at random from those those co-occurring with B will also co-occur with A.

Mutual Information Based Model

Mutual information (MI) allows us to capture the idea that a co-occurrence of low probability

[1] We will consider, for simplicity, similarity between nouns based on the the verbs they co-occur with in the direct object relation but, of course, it would be possible to consider other parts of speech and other relations.

events is more informative than a co-occurrence of high probability events.

In this model, as before, we retain the definitions for feature set membership, TP, precision and recall but again change the association function. Here, the degree of association between a noun n and a verb v is their MI.

$$D_{mi}(n,v) = I(n,v) = \log \frac{P(v,n)}{P(v)P(n)}$$

Accordingly, verb v will be considered to be a feature of noun n if the probability of their co-occurrence is greater than would be expected if verbs and nouns occurred independently.

2.4 Combining Precision and Recall

Although we have defined a pair of numbers for similarity, in applications it will still be necessary to compute a single number in order to determine neighbourhood or cluster membership. There are two obvious ways to optimise a pair of numbers such as precision and recall. The first is to use an arithmetic mean, which optimises the sum of the numbers, and the second is to use a harmonic mean[2], which optimises the product of the numbers.

In an attempt to retain generality, we can allow both alternatives by computing an arithmetic mean of the harmonic mean and the arithmetic mean, noting that the relative importance of each term in an arithmetic mean is controlled by weights (which sum to 1):

$$\begin{aligned} m_h(A,B) &= \left(\frac{2.\mathcal{P}(A,B).\mathcal{R}(A,B)}{\mathcal{P}(A,B)+\mathcal{R}(A,B)}\right) \\ m_a(A,B) &= \beta.\mathcal{P}(A,B) + (1-\beta).\mathcal{R}(A,B) \\ \text{sim}(A,B) &= \gamma.m_h(A,B) + (1-\gamma).m_a(A,B) \end{aligned}$$

where both β and γ lie in the range [0,1]. The resulting similarity sim(A,B) will also lie in the range [0,1] where 0 represents complete lack of similarity and 1 represents equivalence. This formula can be used in combination with any of the models for precision and recall outlined above. Further, the generality allows us to investigate empirically the relative significance of the different terms and thus whether one (or more) might be omitted in future work.

β	γ	Special Case
-	1	harmonic mean
-	0	weighted arithmetic mean
1	0	precision
0	0	recall
0.5	0	unweighted arithmetic mean

Table 1: Table of Special Values of β and γ

Precision and recall can be computed once for every pair of words whereas similarity is something which will be computed for a specific task and will depend on the values of β and γ. Table 1 summarizes some special parameter settings.

3 Empirical Evaluation

In this section, we evaluate the performance of the framework, using the combinatorial and MI-based models of precision and recall, at two application based tasks against Lin's MI-based Measure (sim_{Lin}) and the α-skew Divergence Measure ($\text{sim}_{\alpha\text{sd}}$). The formulae for these measures are given in Figure 1. For the α-skew divergence measure we set $\alpha = 0.99$ since this most closely approximates the Kullback-Leibler divergence measure. The two evaluation tasks used — pseudo-disambiguation and WordNet (Fellbaum, 1998) prediction — are fairly standard for distributional similarity measures. However, in the future we wish to extend our evaluation to other tasks such as malapropism correction (Budanitsky and Hirst, 2001) and PP-attachment ambiguity resolution (Resnik, 1993) and also to the probabilistic model.

Since we use the same data and methodology as in earlier work, some detail is omitted in the subsequent discussion but full details and rationale can be found in Weeds and Weir (2003).

3.1 Pseudo-Disambiguation Task

Pseudo-disambiguation tasks (e.g. Lee, 1999) have become a standard evaluation technique and, in the current context, we may use a word's neighbours to decide which of two co-occurrences is the most likely.

Although pseudo-disambiguation itself is an artificial task, it has relevance in at least two real application areas. First, by replacing occurrences of a particular word in a test suite with a pair or set of words from which a technique must choose, we recreate a simplified version of the word sense disambiguation task; that is,

[2] This is the standard IR measure known as the F-score or F-measure

$$\begin{aligned}
\mathrm{sim}_{\mathrm{Lin}}(n_1,n_2) &= \frac{\sum_{T(n_1)\cap T(n_2)}(I(n_1,v)+I(n_2,v))}{\sum_{T(n_1)} I(n_1,v)+\sum_{T(n_2)} I(n_2,v)} \text{ where } T(n)=\{v : I(n,v>0\} \\
\mathrm{sim}_{\alpha\mathrm{sd}}(n_2,n_1) &= \mathrm{sim}_{\mathrm{KL}}(q||\alpha.r+(1-\alpha).q) \text{ where } q(v)=P(v|n_1), r(v)=P(v|n_2) \\
\mathrm{sim}_{\mathrm{KL}}(q,r) &= \sum_v q(v)\times\log\frac{q(v)}{r(v)} \\
\mathrm{sim}_{\mathrm{Dice}}(n_1,n_2) &= \frac{2.|F(n_1)\cap F(n_2)|}{|F(n_1)|+|F(n_2)|} \text{ where } F(n)=\{v : P(v|n)>0\} \\
\mathrm{sim}_{\mathrm{wn}}(n_1,n_2) &= max_{c_1\in syn(n_1)\wedge c_2\in syn(n_2)}\left(max_{c\in super(c_1)\cap super(c_2)}\frac{2\log P(c)}{\log(P(c_1))+\log(P(c_2))}\right)
\end{aligned}$$

Figure 1: Definitions for similarity measures used throughout this paper

choosing between a fixed number of homonyms based on local context. The second is in language modelling where we wish to estimate the probability of co-occurrences of events but, due to the sparse data problem, it is often the case that a possible co-occurrence has not been seen in the training data.

3.1.1 Methodology

As is common in this field (e.g. Lee, 1999), we study similarity between nouns based on their co-occurrences with verbs in the direct object relation. We study similarity between high and low frequency nouns since we want to investigate any associations between word frequency and quality of neighbours found by the measures but it is impractical to evaluate a large number of similarity measures over all nouns.

2,852,300 lemmatised (noun-verb) direct-object pairs were extracted from the BNC using a shallow parser (Briscoe and Carroll, 1995; Carroll and Briscoe, 1996). From those nouns also occurring in WordNet, we selected the 1000 most frequent[3] nouns and a set of 1000 low frequency[4] nouns.

For each noun, 80% of the available data was randomly selected as training data and the other 20% set aside as test data. Precision and recall were computed for each pair of nouns using the combinatorial and MI models. This data is then available to the application task which will first have to compute the similarity for each pair of nouns based on current parameter settings and select nearest neighbours accordingly.

We converted each noun-verb pair (n, v_1) in the set-aside test data into a noun-verb-verb triple (n, v_1, v_2) where $P(v_1)$ is approximately equal to $P(v_2)$ over all the training data and (n, v_2) has not been seen in the test or training data. A high frequency noun test set and a low frequency noun test set, each containing 10,000 test instances, were then constructed by selecting ten test instances for each noun in a two step process of 1) whilst more than ten triples remained, discarding duplicate triples and 2) randomly selecting ten triples from those remaining after step 1. Each set of test triples was split into five disjoint subsets, containing two triples for each noun, so that average performance and standard error could be computed. Additionally, three of the five subsets were used as a development set to optimise parameters (k, β and γ) and the remaining two used as a test set to find error rates at these optimal settings.

The task is then for the nearest neighbours of noun n to decide which of (n, v_1) and (n, v_2) was the original co-occurrence. Each of n's neighbours, m, is given a vote which is equal to the difference in frequencies of the co-occurrences (m, v_1) and (m, v_2) and which it casts to the co-occurrence in which it appears most frequently. The votes for each co-occurrence are summed over all of the k nearest neighbours of n and the co-occurrence with the most votes wins. Performance is measured as error rate.

$$\text{error} = \frac{1}{T}(\#\text{ of incorrect choices} + \frac{\#\text{ of ties}}{2})$$

where T is the number of test instances.

[3] This corresponds to a frequency range of [576,20561].

[4] We used frequency ranks 3001 to 4000 which correspond to a frequency range of [70,120].

Measure	Noun Frequency			
	high		low	
	params	error	params	error
sim_c	γ=0.25 β=0.8 k=150	0.193	γ=0.25 β=0.75 k=100	0.200
sim_{mi}	γ=0.25 β=0.8 k=170	0.186	γ=0.5 β=0.8 k=120	0.178
$\mathrm{sim}_{\mathrm{Lin}}$	k=50	0.199	k=80	0.186
$\mathrm{sim}_{\alpha\mathrm{sd}}$	k=30	0.233	k=60	0.196

Table 2: Optimal Parameter Settings and Error Rates for the Pseudo-Disambiguation Task

Performance was measured on the development set and the three parameters (β, γ and k) optimised. Of course, it is not possible to make an exhaustive search for the optimal parameter settings, especially when these lie on a continuous scale. In our experiments we tried every combination of parameter settings where β and γ were multiples of 0.1 or 0.25 (from 0 to 1) and k was a multiple of 10 (from 0 to 200) or 50 (from 200 to 1000). These step-sizes gave us smooth results from which we could interpolate intermediate results with reasonable confidence.

3.1.2 Results

Table 2 summarizes the optimal parameter settings (found using the development set) and corresponding mean error rates (in the test set) for the general framework using both the combinatorial (sim_c) and the MI-based (sim_{mi}) models. Results for Lin's MI-based measure ($\mathrm{sim}_{\mathrm{Lin}}$) and the α-skew divergence measure ($\mathrm{sim}_{\alpha\mathrm{sd}}$) are also given and results are divided into those for high frequency nouns and those for low frequency nouns.

Our first observation, based on Table 2, is that the general framework using the MI-based model outperforms the other similarity measures considered for both high and low frequency nouns.

Figure 2 shows how the mean[5] optimal error rate varies with β when $\gamma = 0$. Results are plotted for high and low frequency nouns for both the combinatorial model (sim_c) and the MI-based model (sim_{mi}). From these results, it appears that the pseudo-disambiguation task requires a measure which favours high precision neighbours. By definition, high precision neighbours of noun n are ones which have not occurred with many verbs that do not also occur with n. Such neighbours are unlikely to contribute to many decisions as to which was the original co-occurrence, which is why a higher value of k is required. Our results show that when such neighbours do contribute to the decision, it is likely to be a positive contribution.

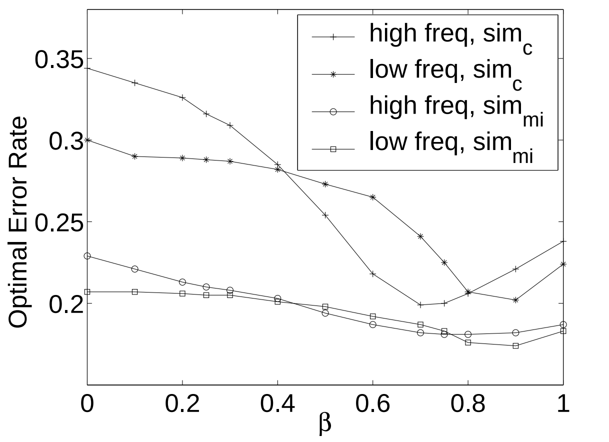

Figure 2: Variation in optimal error rate with β ($\gamma = 0$)

Our last observation is that the setting of γ only ever made a difference of up to 2% in the results obtained, suggesting that the choice between an arithmetic mean and a harmonic mean is less significant than the setting of β, for this task at least. If graphs such as the one in Figure 2 are drawn for other values of γ, we obtain the same pattern but over a smaller range of error rates (since the influence of β decreases as γ increases). In general, the weighted arithmetic mean ($\gamma = 0$) outperforms the unweighted harmonic mean ($\gamma = 1$) by about 1%.

3.2 WordNet Prediction Task

We now turn our attention to the evaluation task used by Lin (1998) – the ability of a similarity measure to predict distance as defined by a measure based on the hyponymy relation in WordNet (version 1.6). This task evaluates the usefulness of a distributional similarity measure as a predictor of semantic similarity and therefore its potential for use in automatic thesaurus generation. An underlying assumption

[5] The mean was taken across the five subsets of the test data. Error bars are not shown but standard errors were of the order of 0.01

is that the hyponymy relation in WordNet is a gold standard for semantic similarity which is, of course, not true. However, we believe that a distributional similarity measure which more closely predicts WordNet, is more likely to be a good predictor of semantic similarity.

3.2.1 Methodology

We will first explain the WordNet-based distance measure (Lin, 1997) and then explain how we determine the similarity between neighbour sets generated using different measures.

The similarity of two nouns in WordNet is defined as the similarity of their maximally similar senses. The *commonality* of two concepts is defined as the maximally specific superclass of those concepts. So, if $syn(n)$ is the set of senses of the noun n in WordNet, $sup(c)$ is the set of (possibly indirect) superclasses of concept c in WordNet and $P(c)$ is the probability that a randomly selected noun refers to an instance of c, then the similarity between n_1 and n_2 can be calculated using the formula for $\mathrm{sim}_{\mathrm{wn}}$ in Figure 1.

The probabilities $P(c)$ are estimated by the frequencies of concepts in SemCor (Miller et al., 1994), a sense-tagged subset of the Brown corpus, noting that the occurrence of a concept refers to instances of all the superclasses of that concept (i.e. $P(\text{root of tree}^6) = 1$).

The k nearest neighbours[7] of each noun, computed using each distributional similarity measure at each parameter setting, are then compared with the k nearest neighbours of the noun according to the WordNet based measure. In order to compute the similarity of two neighbour sets, we transform each neighbour set so that each neighbour is given a rank score of $k-rank$. We do not use the similarity scores directly since these require normalization if different similarity measures (using different scales) are to be compared. Having performed this transformation, the neighbour sets for the same word w may be represented by two ordered sets of words $[w_k, ..., w_1]$ and $[w'_k, ..., w'_1]$. The similarity between such sets is computed using the same calculation as used by Lin (1998) except for simplifications due to the use of ranks:

$$\frac{\sum_{w_i=w'_j} i \times j}{\sum_{i=1}^{k} i^2}$$

where i and j are the rank scores of the words within each neighbour set.

Measure	Noun Frequency			
	high		low	
	params	sim	params	sim
sim_c	$\gamma = 0.25$ $\beta = 0.5$	0.299	$\gamma = 0.5$ $\beta = 0.4$	0.260
sim_{mi}	$\gamma = 0.25$ $\beta = 0.3$	0.317	$\gamma = 0.25$ $\beta = 0.3$	0.274
$\mathrm{sim}_{\mathrm{Lin}}$	-	0.307	-	0.210
$\mathrm{sim}_{\alpha\mathrm{sd}}$	-	0.290	-	0.270

Table 3: Optimal Mean Similarities and Corresponding Parameter Settings Between Thesaurus Entries for WordNet Prediction Task

3.2.2 Results

Table 3 summarizes the optimal mean similarities and parameter settings for the general framework using both the combinatorial (sim_c) and the MI-based (sim_{mi}) models. Results for Lin's MI-based measure ($\mathrm{sim}_{\mathrm{Lin}}$) and the α-skew divergence measure ($\mathrm{sim}_{\alpha\mathrm{sd}}$) are also given and results are divided into those for high frequency nouns and those for low frequency nouns. Standard errors in the optimal mean similarities are not given but were of the order of 0.1.

Our first observation is that the general framework using the MI-based model for precision and recall outperforms all of the other distributional similarity measures.

We also observe that lower values of γ produce better results, particularly for low frequency nouns. For example, when $\gamma = 1$, similarity for low frequency nouns drops to 0.147 using the combinatorial model and 0.177 using the MI-based model.

Third, from Figure 3, it appears that this WordNet prediction task favours measures which select high recall neighbours. Although optimum similarity for the combinatorial model occurs at β=0.5, similarity is always higher for lower values of β than for higher values of β.

[6] The root of the WordNet hyponymy relation is taken to be an imagined superclass of all concepts in WordNet.

[7] As in previous work (Lin, 1998; Weeds and Weir, 2003), we use $k = 200$.

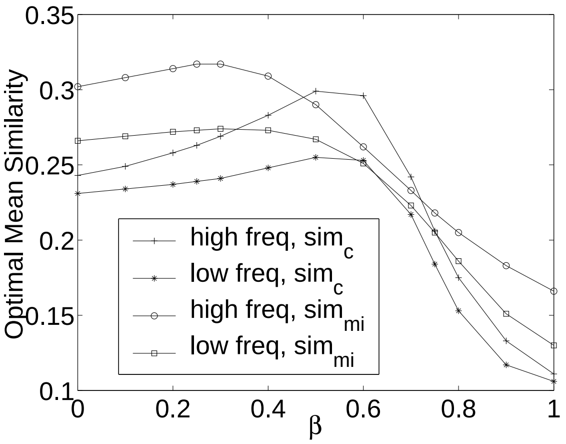

Figure 3: Variation in optimum similarity with β ($\gamma = 0.25$)

4 Characterisations of Existing Measures

In this section, we present some preliminary observations regarding existing similarity measures with reference to the general framework offered in this paper.

4.1 The Dice Coefficient

As shown below, the harmonic mean of precision and recall (or the F-score) using the combinatorial model (Section 2.2) is the Dice Coefficient (See Figure 1 for definition):

$$\begin{aligned}\mathrm{F} &= \frac{2.\mathcal{P}.\mathcal{R}}{\mathcal{P}+\mathcal{R}} = \frac{2.\frac{|TP|}{|F(A)|}.\frac{|TP|}{|F(B)|}}{\frac{|TP|}{|F(A)|}+\frac{|TP|}{|F(B)|}} \\ &= \frac{2.|TP|.|TP|}{|TP|.|F(A)|+|TP|.|F(B)|} \\ &= \frac{2.|TP|}{|F(A)|+|F(B)|} = \mathrm{sim}_{\mathrm{Dice}}(A,B)\end{aligned}$$

Accordingly, when γ is set to 1 and the combinatorial model is used, our framework reduces to the Dice Coefficient.

4.2 Lin's MI-based Measure

There are parallels between Lin's MI-Based Measure and the Dice Coefficient in that both measures compute a ratio between what is shared by the descriptions of both nouns and the sum of the descriptions of each noun. The relationship between Lin's Measure and the F-Score (using the MI-based model (Section 2.3)) is shown below:

$$\mathrm{F} = \frac{2.\mathcal{P}.\mathcal{R}}{\mathcal{P}+\mathcal{R}} = \frac{2.\frac{\sum_{TP} I(A,v)}{\sum_{F(A)} I(A,v)}.\frac{\sum_{TP} I(B,v)}{\sum_{F(B)} I(B,v)}}{\frac{\sum_{TP} I(A,v)}{\sum_{F(A)} I(A,v)}+\frac{\sum_{TP} I(B,v)}{\sum_{F(B)} I(B,v)}}$$

$$= \frac{2.\sum_{TP} I(A,v).\sum_{TP} I(B,v)}{\sum_{TP} I(B,v).\sum_{F(A)} I(A,v)+\sum_{TP} I(A,v).\sum_{F(B)} I(B,v)}$$

Now, if $\sum_{TP} I(A,v) = \sum_{TP} I(B,v)$, it follows:

$$\mathrm{F} = \frac{2.\sum_{TP} I(A,v)}{\sum_{F(A)} I(A,v)+\sum_{F(B)} I(B,v)}$$

$$= \frac{\sum_{TP} I(A,v)+I(B,v)}{\sum_{F(A)} I(A,v)+\sum_{F(B)} I(B,v)} = \mathrm{sim}_{\mathrm{Lin}}(A,B)$$

Thus, when the MI-based model is used, $\gamma = 1$ and the condition $\sum_{TP} I(A,v) = \sum_{TP} I(B,v)$ holds, our framework reduces to Lin's Measure. Further, by considering the definition for MI, we see that the necessary condition for equivalence holds when nouns A and B have exactly the same probability of co-occurring with each of their shared verbs. As the probabilities diverge, so will the similarity between the neighbours computed using $\gamma = 1$ and the neighbours computed using Lin's Measure.

In order to investigate how good an approximation the F-Score is to Lin's Measure when the condition does not hold, we compared the neighbours according to each measure using the neighbour set comparison technique outlined in Section 3.2.1. At $\gamma = 1$, the similarity was 0.967 for high frequency nouns and 0.922 for low frequency nouns. This is much higher than similarities between other standard distributional similarity measures. For example, the similarity between Lin's Measure and the α-skew divergence measure is 0.591 for high frequency nouns and 0.360 for low frequency nouns. Interestingly, however, the optimal approximation of Lin's Measure was found using $\gamma = 0.75$ and $\beta = 0.5$. With these settings, the similarity was 0.987 for high frequency nouns and 0.977 for low frequency nouns. This suggests that Lin's Measure allows more compensation for lack of recall by precision and vice versa than the F-Score.

4.3 The α-skew Divergence Measure

From examination of its definition and consideration of its results on the two evaluation tasks, we predicted that the α-skew divergence measure might be approximated by using a lower value of β (i.e. high recall). This was supported by comparisons of the neighbour sets found using the α-skew divergence measure and those found using the MI-Based model. Optimal similarity (0.760 and 0.725 respectively) was found at $\gamma = 0.0$ and $\beta = 0.0$ for high frequency nouns and at $\gamma = 0.25$ and $\beta = 0.0$ for low frequency nouns. Further, similarity between the measures drops rapidly once β rises above 0.3.

5 Conclusions and Further Work

Using the MI-based model for precision and recall and with a parameter setting of $\gamma = 1.0$, the general framework for distributional similarity proposed herein closely approximates Lin's (1998) Measure. However, we have shown that using a much lower value of γ so that the combination of precision and recall is closer to a weighted arithmetic mean than a harmonic mean yields better results in the two application tasks considered here. This is because the relative importance of precision and recall can be tuned to the task at hand.

Further, we have shown that pseudo-disambiguation is a task which requires high precision neighbours whereas WordNet prediction is a task which requires high recall neighbours. Accordingly, it is not clear how a single (unparameterised) similarity measure could give optimum results on both tasks.

In the future, we intend to extend the work to the characterisation of other tasks and other existing similarity measures. As well as their, usually implicit, use of precision and recall, the main difference between existing similarity measures will be the models in which precision and recall are defined. We have explored two such models here – a combinatorial model and a MI-based model – and have shown that the MI-based model achieves significantly improved results over the combinatorial model. We propose to investigate other models such as the probabilistic one given in Section 2.3.

6 Acknowledgements

We would like to thank John Carroll for the use of his parser, Adam Kilgarriff and Bill Keller for valuable discussions and the UK EPSRC for its studentship to the first author.

References

E. Briscoe and J. Carroll. 1995. Developing and evaluating a probabilistic lr parser of part-of-speech and punctuation labels. In *4th ACL/SIGDAT International Workshop on Parsing Technologies*, pages 48–58.

A. Budanitsky and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *NAACL-01.*

J. Carroll and E. Briscoe. 1996. Apportioning development effort in a probabilistic lr parsing system through evaluation. In *ACL/SIGDAT Conference on Empirical Methods in Natural Language Processing*, pages 92–100.

T.M. Cover and J.A. Thomas. 1991. *Elements of Information Theory.* Wiley, New York.

U. Essen and V. Steinbiss. 1992. Cooccurrence smoothing for stochastic language modelling. *ICASSP 92*, 1:161–164.

C. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

W.B. Frakes and R. Baeza-Yates, editors. 1992. *Information Retrieval, Data Structures and Algorithms.* Prentice Hall.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *ACL-90*, pages 268–275.

L. Lee. 1999. Measures of distributional similarity. In *ACL-99.*

B. Levin. 1993. *Towards a Lexical Organization of English Verbs.* Chicago University Press.

D. Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of ACL/EACL-97*, pages 64–71.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL '98.*

G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas. 1994. Using a semantic concordance for sense identification. In *ARPA Human Language Technology Workshop.*

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of similar words. In *ACL'93.*

C. Radhakrishna Rao. 1983. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankyha: The Indian Journal of Statistics*, 44(A):1–22.

P. Resnik. 1993. *Selection and Information: A Class-Based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania.

G. Salton and M.J. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.

H. Silber and K. McCoy. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics*, 28(4).

J. Weeds and D. Weir. 2003. Finding and evaluating sets of nearest neighbours. In *Proceedings of 2nd Conference on Corpus Linguistics.*