Characterising Measures of Lexical Distributional Similarity

Julie Weeds, David Weir and Diana McCarthy
Department of Informatics
University of Sussex
Brighton, BN1 9QH, UK
{juliewe, davidw,dianam}@sussex.ac.uk

Abstract

This work investigates the variation in a word's distributionally nearest neighbours with respect to the similarity measure used. We identify one type of variation as being the relative frequency of the neighbour words with respect to the frequency of the target word. We then demonstrate a three-way connection between relative frequency of similar words, a concept of distributional gnerality and the semantic relation of hyponymy. Finally, we consider the impact that this has on one application of distributional similarity methods (judging the compositionality of collocations).

1 Introduction

Over recent years, many Natural Language Processing (NLP) techniques have been developed that might benefit from knowledge of distributionally similar words, i.e., words that occur in similar contexts. For example, the sparse data problem can make it difficult to construct language models which predict combinations of lexical events. Similarity-based smoothing (Brown et al., 1992; Dagan et al., 1999) is an intuitively appealing approach to this problem where probabilities of unseen co-occurrences are estimated from probabilities of seen co-occurrences of distributionally similar events.

Other potential applications apply the hypothesised relationship (Harris, 1968) between distributional similarity and semantic similarity; i.e., similarity in the meaning of words can be predicted from their distributional similarity. One advantage of automatically generated thesauruses (Grefenstette, 1994; Lin, 1998; Curran and Moens, 2002) over large-scale manually created thesauruses such as WordNet (Fellbaum, 1998) is that they might be tailored to a particular genre or domain.

However, due to the lack of a tight definition for the concept of distributional similarity and the broad range of potential applications, a large number of measures of distributional similarity have been proposed or adopted (see Section 2). Previous work on the evaluation of distributional similarity methods tends to either compare sets of distributionally similar words to a manually created semantic resource (Lin, 1998; Curran and Moens, 2002) or be oriented towards a particular task such as language modelling (Dagan et al., 1999; Lee, 1999). The first approach is not ideal since it assumes that the goal of distributional similarity methods is to predict semantic similarity and that the semantic resource used is a valid gold standard. Further, the second approach is clearly advantageous when one wishes to apply distributional similarity methods in a particular application area. However, it is not at all obvious that one universally best measure exists for all applications (Weeds and Weir, 2003). Thus, applying a distributional similarity technique to a new application necessitates evaluating a large number of distributional similarity measures in addition to evaluating the new model or algorithm.

We propose a shift in focus from attempting to discover the overall best distributional similarity measure to analysing the statistical and linguistic properties of sets of distributionally similar words returned by different measures. This will make it possible to predict in advance of any experimental evaluation which distributional similarity measures might be most appropriate for a particular application.

Further, we explore a problem faced by the automatic thesaurus generation community, which is that distributional similarity methods do not seem to offer any obvious way to distinguish between the semantic relations of synonymy, antonymy and hyponymy. Previous work on this problem (Caraballo, 1999; Lin et al., 2003) involves identifying specific phrasal patterns within text e.g., "Xs and other Ys" is used as evidence that X is a hyponym of Y. Our work explores the connection between relative frequency, distributional generality and semantic generality with promising results.

The rest of this paper is organised as follows. In Section 2, we present ten distributional similarity measures that have been proposed for use in NLP. In Section 3, we analyse the variation in neighbour sets returned by these measures. In Section 4, we take one fundamental statistical property (word frequency) and analyse correlation between this and the nearest neighbour sets generated. In Section 5, we relate relative frequency to a concept of distributional generality and the semantic relation of hyponymy. In Section 6, we consider the effects that this has on a potential application of distributional similarity techniques, which is judging compositionality of collocations.

2 Distributional similarity measures

In this section, we introduce some basic concepts and then discuss the ten distributional similarity measures used in this study.

The *co-occurrence types* of a target word are the contexts, c, in which it occurs and these have associated frequencies which may be used to form probability estimates. In our work, the co-occurrence types are always grammatical dependency relations. For example, in Sections 3 to 5, similarity between nouns is derived from their co-occurrences with verbs in the direct-object position. In Section 6, similarity between verbs is derived from their subjects and objects. The k nearest neighbours of a target word w are the k words for which similarity with w is greatest. Our use of the term similarity measure encompasses measures which should strictly be referred to as *distance*, *divergence* or *dissimilarity* measures. An increase in distance correlates with a decrease in similarity. However, either type of measure can be used to find the k nearest neighbours of a target word.

Table 1 lists ten distributional similarity measures. The cosine measure (Salton and McGill, 1983) returns the cosine of the angle between two vectors.

The Jensen-Shannon (JS) divergence measure (Rao, 1983) and the α-skew divergence measure (Lee, 1999) are based on the Kullback-Leibler (KL) divergence measure. The KL divergence, or relative entropy, $D(p||q)$, between two probability distribution functions p and q is defined (Cover and Thomas, 1991) as the “inefficiency of assuming that the distribution is q when the true distribution is p”: $D(p||q) = \sum_c p \log \frac{p}{q}$. However, $D(p||q) = \infty$ if there are any contexts c for which $p(c) > 0$ and $q(c) = 0$. Thus, this measure cannot be used directly on maximum likelihood estimate (MLE) probabilities. One possible solution is to use the JS divergence measure, which measures the cost of using the average distribution in place of each individual distribution. Another is the α-skew divergence measure, which uses the p distribution to smooth the q distribution. The value of the parameter α controls the extent to which the KL divergence is approximated. We use $\alpha = 0.99$ since this provides a close approximation to the KL divergence and has been shown to provide good results in previous research (Lee, 2001).

The confusion probability (Sugawara et al., 1985) is an estimate of the probability that one word can be substituted for another. Words w_1 and w_2 are completely *confusable* if we are equally as likely to see w_2 in a given context as we are to see w_1 in that context.

Jaccard's coefficient (Salton and McGill, 1983) calculates the proportion of features belonging to either word that are shared by both words. In the simplest case, the features of a word are defined as the contexts in which it has been seen to occur. sim_{ja+mi} is a variant (Lin, 1998) in which the features of a word are those contexts for which the pointwise mutual information (MI) between the word and the context is positive, where MI can be calculated using $I(c, w) = \log \frac{P(c|w)}{P(c)}$. The related Dice Coefficient (Frakes and Baeza-Yates, 1992) is omitted here since it has been shown (van Rijsbergen, 1979) that Dice and Jaccard's Coefficients are monotonic in each other.

Lin's Measure (Lin, 1998) is based on his information-theoretic similarity theorem, which states, “the similarity between A and B is measured by the ratio between the amount of information needed to state the commonality of A and B and the information needed to fully describe what A and B are.”

The final three measures are settings in the additive MI-based Co-occurrence Retrieval Model (AMCRM) (Weeds and Weir, 2003; Weeds, 2003). We can measure the *precision* and the *recall* of a potential neighbour's retrieval of the co-occurrences of the target word, where the sets of required and retrieved co-occurrences ($F(w_1)$ and $F(w_2)$ respectively) are those co-occurrences for which MI is positive. Neighbours with both high precision and high recall retrieval can be obtained by computing

Measure	Function
cosine	$sim_{cm}(w_2, w_1) = \frac{\sum_c P(c\mid w_1).P(c\mid w_2)}{\sqrt{\sum_c P(c\mid w_1)^2 \sum_c P(c\mid w_2)^2}}$
Jens.-Shan.	$dist_{js}(w_2, w_1) = \frac{1}{2}\left(D\left(p\Vert\frac{p+q}{2}\right) + D\left(q\Vert\frac{p+q}{2}\right)\right)$ where $p = P(c\mid w_1)$ and $q = P(c\mid w_2)$
α-skew	$dist_\alpha(w_2, w_1) = D\left(p\Vert(\alpha.q + (1-\alpha).p)\right)$ where $p = P(c\mid w_1)$ and $q = P(c\mid w_2)$
conf. prob.	$sim_{cp}(w_2\mid w_1) = \sum_c \frac{P(w_1\mid c).P(w_2\mid c).P(c)}{P(w_1)}$
Jaccard's	$sim_{ja}(w_2, w_1) = \frac{\lvert F(w_1)\cap F(w_2)\rvert}{\lvert F(w_1)\cup F(w_2)\rvert}$where $F(w) = \{c : P(c\mid v) > 0\}$
Jacc.+MI	$sim_{ja+mi}(w_2, W_1) = \frac{\lvert F(w_1)\cap F(w_2)\rvert}{\lvert F(w_1)\cup F(w_2)\rvert}$ where $F(w) = \{c : I(c, w) > 0\}$
Lin's	$sim_{lin}(w_2, w_1) = \frac{\sum_{F(w_1)\cap F(w_2)}(I(c,w_1)+I(c,w_2))}{\sum_{F(w_1)} I(c,w_1)+\sum_{F(w_2)} I(c,w_2)}$ where $F(w) = \{c : I(c, w) > 0\}$
precision	$sim_{\mathcal{P}}(w_2, w_1) = \frac{\sum_{F(w_1)\cap F(w_2)} I(c,w_2)}{\sum_{F(w_2)} I(c,w_2)}$ where $F(w) = \{c : I(c, w) > 0\}$
recall	$sim_{\mathcal{R}}(w_2, w_1) = \frac{\sum_{F(w_1)\cap F(w_2)} I(c,w_1)}{\sum_{F(w_1)} I(c,w_1)}$ where $F(w) = \{c : I(c, w) > 0\}$
harm. mean	$sim_{hm}(w_2, w_1) = \frac{2.sim_{\mathcal{P}}(w_2,w_1).sim_{\mathcal{R}(w_2,w_1)}}{sim_{\mathcal{P}}(w_2,w_1)+sim_{\mathcal{R}}(w_2,w_1)}$ where $F(w) = \{c : I(c, w) > 0\}$

Table 1: Ten distributional similarity measures

their *harmonic mean* (or F-score).

3 Overlap of neighbour sets

We have described a number of ways of calculating distributional similarity. We now consider whether there is substantial variation in a word's distributionally nearest neighbours according to the chosen measure. We do this by calculating the overlap between neighbour sets for 2000 nouns generated using different measures from direct-object data extracted from the British National Corpus (BNC).

3.1 Experimental set-up

The data from which sets of nearest neighbours are derived is direct-object data for 2000 nouns extracted from the BNC using a robust accurate statistical parser (RASP) (Briscoe and Carroll, 2002). For reasons of computational efficiency, we limit ourselves to 2000 nouns and direct-object relation data. Given the goal of comparing neighbour sets generated by different measures, we would not expect these restrictions to affect our findings. The complete set of 2000 nouns (WS$_{comp}$) is the union of two sets WS$_{high}$ and WS$_{low}$ for which nouns were selected on the basis of frequency: WS$_{high}$ contains the 1000 most frequently occurring nouns (frequency $>$ 500), and WS$_{low}$ contains the nouns ranked 3001-4000 (frequency $\approx$ 100). By excluding mid-frequency nouns, we obtain a clear separation between high and low frequency nouns. The complete data-set consists of 1,596,798 co-occurrence tokens distributed over 331,079 co-occurrence types. From this data, we computed the similarity between every pair of nouns according to each distributional similarity measure. We then generated ranked sets of nearest neighbours (of size $k = 200$ and where a word is excluded from being a neighbour of itself) for each word and each measure.

For a given word, we compute the overlap between neighbour sets using a comparison technique adapted from Lin (1998). Given a word w, each word w' in WS$_{comp}$ is assigned a `rank score` of $k - rank$ if it is one of the k nearest neighbours of w using measure m and zero otherwise. If NS(w, m) is the vector of such scores for word w and measure m, then the overlap, $C($NS$(w, m_1),$ NS$(w, m_2))$, of two neighbour sets is the cosine between the two vectors: $C($NS$(w, m_1),$ NS$(w, m_2)) =$

$$\frac{\sum_{w'} r_{m_1}(w', w) \times r_{m_2}(w', w)}{\sum_{i=1}^{k} i^2}$$

The overlap score indicates the extent to which sets share members and the extent to which they are in the same order. To achieve an overlap score of 1, the sets must contain exactly the same items in exactly the same order. An overlap score of 0 is obtained if the sets do not contain any common items. If two sets share roughly half their items and these shared items are dispersed throughout the sets in a roughly similar order, we would expect the overlap between sets to be around 0.5.

	cm	js	α	cp	ja	$ja+mi$	lin
cm	1.0(0.0)	**0.69**(0.12)	0.53(0.15)	0.33(0.09)	0.26(0.12)	0.28(0.15)	0.32(0.15)
js	0.69(0.12)	1.0(0.0)	**0.81**(0.10)	0.46(0.31)	0.48(0.18)	0.49(0.20)	0.55(0.16)
α	0.53(0.15)	**0.81**(0.10)	1.0(0.0)	0.61(0.08)	0.4(0.27)	0.39(0.25)	0.48(0.19)
cp	0.33(0.09)	0.46(0.31)	**0.61**(0.08)	1.0(0.0)	0.24(0.24)	0.20(0.18)	0.29(0.15)
ja	0.26(0.12)	0.48(0.18)	0.4(0.27)	0.24(0.24)	1.0(0.0)	**0.81**(0.08)	0.69(0.09)
$ja+mi$	0.28(0.15)	0.49(0.20)	0.39(0.25)	0.20(0.18)	**0.81**(0.08)	1.0(0.0)	0.81(0.10)
lin	0.32(0.15)	0.55(0.16)	0.48(0.19)	0.29(0.15)	0.69(0.09)	**0.81**(0.10)	1.0(0.0)

Table 2: Cross-comparison of first seven similarity measures in terms of mean overlap of neighbour sets and corresponding standard deviations.

	$\mathcal{P}$	$\mathcal{R}$	hm
cm	0.18(0.10)	**0.31**(0.13)	0.30(0.14)
js	0.19(0.12)	**0.55**(0.18)	0.51(0.18)
α	0.08(0.08)	**0.74**(0.14)	0.41(0.23)
cp	0.03(0.04)	**0.57**(0.10)	0.25(0.18)
ja	0.36(0.30)	0.38(0.30)	**0.74**(0.14)
$ja+mi$	0.42(0.30)	0.40(0.31)	**0.86**(0.07)
lin	0.46(0.25)	0.52(0.22)	**0.95**(0.039)

Table 3: Mean overlap scores for seven similarity measures with precision, recall and the harmonic mean in the AMCRM.

3.2 Results

Table 2 shows the mean overlap score between every pair of the first seven measures in Table 1 calculated over WS_{comp}. Table 3 shows the mean overlap score between each of these measures and precision, recall and the harmonic mean in the AMCRM. In both tables, standard deviations are given in brackets and boldface denotes the highest levels of overlap for each measure. For compactness, each measure is denoted by its subscript from Table 1.

Although overlap between most pairs of measures is greater than expected if sets of 200 neighbours were generated randomly from WS_{comp} (in this case, average overlap would be 0.08 and only the overlap between the pairs (α,$\mathcal{P}$) and (cp,$\mathcal{P}$) is not significantly greater than this at the 1% level), there are substantial differences between the neighbour sets generated by different measures. For example, for many pairs, neighbour sets do not appear to have even half their members in common.

4 Frequency analysis

We have seen that there is a large variation in neighbours selected by different similarity measures. In this section, we analyse how neighbour sets vary with respect to one fundamental statistical property — word frequency. To do this, we measure the bias in neighbour sets towards high frequency nouns and consider how this varies depending on whether the target noun is itself a high frequency noun or low frequency noun.

4.1 Measuring bias

If a measure is biased towards selecting high frequency words as neighbours, then we would expect that neighbour sets for this measure would be made up mainly of words from WS_{high}. Further, the more biased the measure is, the more highly ranked these high frequency words will tend to be. In other words, there will be high overlap between neighbour sets generated considering all 2000 nouns as potential neighbours and neighbour sets generated considering just the nouns in WS_{high} as potential neighbours. In the extreme case, where all of a noun's k nearest neighbours are high frequency nouns, the overlap with the high frequency noun neighbour set will be 1 and the overlap with the low frequency noun neighbour set will be 0. The inverse is, of course, true if a measure is biased towards selecting low frequency words as neighbours.

If $\text{NS}_{wordset}$ is the vector of neighbours (and associated rank scores) for a given word, w, and similarity measure, m, and generated considering just the words in *wordset* as potential neighbours, then the overlap between two neighbour sets can be computed using a cosine (as before). If $C_{high} = C(\text{NS}_{comp}, \text{NS}_{high})$ and $C_{low} = C(\text{NS}_{comp}, \text{NS}_{low})$, then we compute the bias towards high frequency neighbours for word w using measure m as: $biashigh_m(w) = \frac{C_{high}}{C_{high}+C_{low}}$

The value of this normalised score lies in the range [0,1] where 1 indicates a neighbour set completely made up of high frequency words, 0 indicates a neighbour set completely made up of low frequency words and 0.5 indicates a neighbour set with no biases towards high or low frequency words. This score is more informative than simply calculating the proportion of high

	high freq. target nouns	**low** freq. target nouns
cm	0.90	0.87
js	0.94	0.70
α	0.98	0.90
cp	1.00	0.99
ja	0.99	0.21
$ja + mi$	0.95	0.14
lin	0.85	0.38
$\mathcal{P}$	0.12	0.04
$\mathcal{R}$	0.99	0.98
hm	0.92	0.28

Table 4: Mean value of *biashigh* according to measure and frequency of target noun.

and low frequency words in each neighbour set because it weights the importance of neighbours by their rank in the set. Thus, a large number of high frequency words in the positions closest to the target word is considered more biased than a large number of high frequency words distributed throughout the neighbour set.

4.2 Results

Table 4 shows the mean value of the *biashigh* score for every measure calculated over the set of high frequency nouns and over the set of low frequency nouns. The standard deviations (not shown) all lie in the range [0,0.2]. Any deviation from 0.5 of greater than 0.0234 is significant at the 1% level.

For all measures and both sets of target nouns, there appear to be strong tendencies to select neighbours of particular frequencies. Further, there appears to be three classes of measures: those that select high frequency nouns as neighbours regardless of the frequency of the target noun (cm, js, α, cp and $\mathcal{R}$); those that select low frequency nouns as neighbours regardless of the frequency of the target noun ($\mathcal{P}$); and those that select nouns of a similar frequency to the target noun (ja, $ja + mi$, lin and hm).

This can also be considered in terms of *distributional generality*. By definition, recall prefers words that have occurred in more of the contexts that the target noun has, regardless of whether it occurs in other contexts as well i.e., it prefers distributionally more general words. The probability of this being the case increases as the frequency of the potential neighbour increases and so, recall tends to select high frequency words. In contrast, precision prefers words that have occurred in very few contexts that the target word has not i.e., it prefers distributionally more specific words. The probability of this being the case increases as the frequency of the potential neighbour decreases and so, precision tends to select low frequency words. The harmonic mean of precision and recall prefers words that have both high precision and high recall. The probability of this being the case is highest when the words are of similar frequency and so, the harmonic mean will tend to select words of a similar frequency.

5 Relative frequency and hyponymy

In this section, we consider the observed frequency effects from a semantic perspective.

The concept of *distributional generality* introduced in the previous section has parallels with the linguistic relation of hyponymy, where a hypernym is a semantically more general term and a hyponym is a semantically more specific term. For example, `animal` is an (indirect[1]) hypernym of `dog` and conversely `dog` is an (indirect) hyponym of `animal`. Although one can obviously think of counter-examples, we would generally expect that the more specific term `dog` can only be used in contexts where `animal` can be used and that the more general term `animal` might be used in all of the contexts where `dog` is used and possibly others. Thus, we might expect that distributional generality is correlated with semantic generality — a word has high recall/low precision retrieval of its hyponyms' co-occurrences and high precision/low recall retrieval of its hypernyms' co-occurrences.

Thus, if n_1 and n_2 are related and $\mathcal{P}(n_2, n_1) > \mathcal{R}(n_2, n_1)$, we might expect that n_2 is a hyponym of n_1 and vice versa. However, having discussed a connection between frequency and distributional generality, we might also expect to find that the frequency of the hypernymic term is greater than that of the hyponymic term. In order to test these hypotheses, we extracted all of the possible hyponym-hypernym pairs (20, 415 pairs in total) from our list of 2000 nouns (using WordNet 1.6). We then calculated the proportion for which the direction of the hyponymy relation could be accurately predicted by the relative values of precision and recall and the proportion for which the direction of the hyponymy relation could be accurately predicted by relative frequency. We found that the direction of the hyponymy relation is correlated in the predicted direction with the precision-recall

[1] There may be other concepts in the hypernym chain between `dog` and `animal` e.g. `carnivore` and `mammal`.

values in 71% of cases and correlated in the predicted direction with relative frequency in 70% of cases. This supports the idea of a three-way linking between distributional generality, relative frequency and semantic generality. We now consider the impact that this has on a potential application of distributional similarity methods.

6 Compositionality of collocations

In its most general sense, a collocation is a habitual or lexicalised word combination. However, some collocations such as *strong tea* are compositional, i.e., their meaning can be determined from their constituents, whereas others such as *hot dog* are not. Both types are important in language generation since a system must choose between alternatives but only non-compositional ones are of interest in language understanding since only these collocations need to be listed in the dictionary.

Baldwin et al. (2003) explore empirical models of compositionality for noun-noun compounds and verb-particle constructions. Based on the observation (Haspelmath, 2002) that compositional collocations tend to be hyponyms of their head constituent, they propose a model which considers the semantic similarity between a collocation and its constituent words.

McCarthy et al. (2003) also investigate several tests for compositionality including one (`simplexscore`) based on the observation that compositional collocations tend to be similar in meaning to their constituent parts. They extract co-occurrence data for 111 phrasal verbs (e.g. *rip off*) and their simplex constituents (e.g. *rip*) from the BNC using RASP and calculate the value of sim_{lin} between each phrasal verb and its simplex constituent. The test `simplexscore` is used to rank the phrasal verbs according to their similarity with their simplex constituent. This ranking is correlated with human judgements of the compositionality of the phrasal verbs using Spearman's rank correlation coefficient. The value obtained (0.0525) is disappointing since it is not statistically significant (the probability of this value under the null hypothesis of "no correlation" is 0.3).[2]

However, Haspelmath (2002) notes that a compositional collocation is not just similar to one of its constituents — it can be considered to be a hyponym of its head constituent. For example, "strong tea" is a type of "tea" and "to rip up" is a way of "ripping".

Measure	r_s	$P(r_s)$ under H_0
sim_{lin}	0.0525	0.2946
precision	-0.160	0.0475
recall	**0.219**	**0.0110**
harmonic mean	0.011	0.4562

Table 5: Correlation with compositionality for different similarity measures

Thus, we hypothesised that a distributional measure which tends to select more general terms as neighbours of the phrasal verb (e.g. recall) would do better than measures that tend to select more specific terms (e.g. precision) or measures that tend to select terms of a similar specificity (e.g sim_{lin} or the harmonic mean of precision and recall).

Table 5 shows the results of using different similarity measures with the `simplexscore` test and data of McCarthy et al. (2003). We now see significant correlation between compositionality judgements and distributional similarity of the phrasal verb and its head constituent. The correlation using the recall measure is significant at the 5% level; thus we can conclude that if the simplex verb has high recall retrieval of the phrasal verb's co-occurrences, then the phrasal is likely to be compositional. The correlation score using the precision measure is negative since we would not expect the simplex verb to be a hyponym of the phrasal verb and thus, if the simplex verb does have high precision retrieval of the phrasal verb's co-occurrences, it is less likely to be compositional.

Finally, we obtained a very similar result (0.217) by ranking phrasals according to their inverse relative frequency with their simplex constituent (i.e., $\frac{freq(simplex)}{freq(phrasal)}$). Thus, it would seem that the three-way connection between distributional generality, hyponymy and relative frequency exists for verbs as well as nouns.

7 Conclusions and further work

We have presented an analysis of a set of distributional similarity measures. We have seen that there is a large amount of variation in the neighbours selected by different measures and therefore the choice of measure in a given application is likely to be important.

We also identified one of the major axes of variation in neighbour sets as being the frequency of the neighbours selected relative to the frequency of the target word. There are three

[2] Other tests for compositionality investigated by McCarthy et al. (2003) do much better.

major classes of distributional similarity measures which can be characterised as 1) higher frequency selecting or high recall measures; 2) lower frequency selecting or high precision measures; and 3) similar frequency selecting or high precision and recall measures.

A word tends to have high recall similarity with its hyponyms and high precision similarity with its hypernyms. Further, in the majority of cases, it tends to be more frequent than its hyponyms and less frequent than its hypernyms. Thus, there would seem to a three way correlation between word frequency, distributional generality and semantic generality.

We have considered the impact of these observations on a technique which uses a distributional similarity measure to determine compositionality of collocations. We saw that in this application we achieve significantly better results using a measure that tends to select higher frequency words as neighbours rather than a measure that tends to select neighbours of a similar frequency to the target word.

There are a variety of ways in which this work might be extended. First, we could use the observations about distributional generality and relative frequency to aid the process of organising distributionally similar words into hierarchies. Second, we could consider the impact of frequency characteristics in other applications. Third, for the general application of distributional similarity measures, it would be useful to find other characteristics by which distributional similarity measures might be classified.

Acknowledgements

This work was funded by a UK EPSRC studentship to the first author, UK EPSRC project GR/S26408/01 "Natural Language Service Composition (NatHab)" and UK EPSRC project GR/N36494/01 "Robust Accurate Statistical Parsing (RASP)".

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions*, pages 89–96, Sapporo, Japan.

Edward Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of LREC-2002*, pages 1499–1504.

P.F. Brown, V.J. DellaPietra, P.V deSouza, J.C. Lai, and R.L. Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Sharon Caraballo. 1999. Automatic construction of a hypernym-labelled noun hierarchy from text. In *Proceedings of ACL-99*, pages 120–126.

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

James R. Curran and Marc Moens. 2002. Improvements in automatic thesaurus extraction. In *ACL-SIGLEX Workshop on Unsupervised Lexical Acquisition*, Philadelphia.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning Journal*, 34.

Christiane 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.

Gregory Grefenstette. 1994. Corpus-derived first-, second- and third-order word affinities. In *Proceedings of Euralex*, pages 279–290, Amsterdam.

Zelig S. Harris. 1968. *Mathematical Structures of Language*. Wiley, New York.

Martin Haspelmath. 2002. *Understanding Morphology*. Arnold Publishers.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of ACL-1999*, pages 23–32.

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

Dekang Lin, Shaojun Zhao, Lijuan Qin, and Ming Zhou. 2003. Identifying synonyms among distributionally similar words. In *Proceedings of IJCAI-03*, pages 1492–1493.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL '98*, pages 768–774, Montreal.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL-2003 Workshop on Multiword Expressions*, pages 73–80, Sapporo, Japan.

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

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

K.M. Sugawara, K. Nishimura, K. Toshioka, M. Okachi, and T. Kaneko. 1985. Isolated word recognition using hidden markov models. In *Proceedings of the ICASSP-1985*, pages 1–4.

C.J. van Rijsbergen. 1979. *Information Retrieval*. Butterworths, second edition.

Julie Weeds and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP-2003*, pages 81–88, Sapporo, Japan.

Julie Weeds. 2003. *Measures and Applications of Lexical Distributional Similarity*. Ph.D. thesis, Department of Informatics, University of Sussex.