Asymmetric Association Measures

Lukas Michelbacher
Institute for NLP
Universität Stuttgart
michells@ifnlp.org

Stefan Evert
Cognitive Science
Universität Osnabrück
stefan.evert@uos.de

Hinrich Schütze
Institute for NLP
Universität Stuttgart
hs999@ifnlp.org

Abstract

Human word associations are asymmetric or directed. When hearing a word like *mango*, *fruit* is one of the first associations that come to mind. But when hearing *fruit*, we are more likely to come up with common fruits like *apple* or *orange* than the less frequent *mango*. Similar asymmetry effects have been observed for collocations, recurrent syntagmatic word combinations that are often lexically determined. Despite these intuitions, virtually all corpus-based measures of the statistical association between words are symmetric. In this paper, we propose two asymmetric, directed association measures, viz. conditional probability and a rank measure derived from the χ^2 test. The goal of this paper is to determine to what extent these two measures of directed "corpus association" can be used as a model for directed "psychological association" in the human mind. Both measures were implemented and applied to a large data set, the *British National Corpus* (BNC). The results were evaluated against directed human association data obtained from the *University of South Florida (USF) Free Association Norms* database. We find that the new measures are able to distinguish between highly symmetric and highly asymmetric pairs to some extent, but the overall accuracy in predicting the degree of asymmetry is low.

Keywords

Association measures, collocations

1 Introduction

Statistical association measures (see e.g. [2]) are commonly used to quantify collocational strength [3], i.e. the tendency of words such as *day* and *night* to "keep each other company". Although Firth speaks of a *mutual expectancy* between collocates [3] and virtually all association measures are symmetric (i.e. the calculated scores do not depend on the order in which the two words are given), native speakers have strong intuitions that in many cases one term in a collocation is more "important" for the other than vice versa.

Several authors discuss this phenomenon. Sinclair distinguishes between *upward* and *downward collocation* based on the occurrence frequencies of the two collocates [17, p. xxiii]. In a similar vein, Kjellmer [7] distinguishes three kinds of collocations: (i) *right and left predictive* collocations like *aurora borealis*, where the first word suggests the second as auch as the second suggests the first; (ii) *right predictive* collocations, in which the first word suggest the second but not vice versa (e.g. *wellington boots*); and (iii) *left predictive* collocations like *arms akimbo*, where the second word suggests the first but not the other way around. Hausmann's definition of collocations [5], which focuses on learner dictionaries, distinguishes between *base* and *collocate*. The base of a collocation, typically a noun, retains its regular meaning whereas the collocate is lexically determined and its meaning is modified or weakened. A classic example is *heavy smoker* with base *smoker*. The relation between the two words in such a collocation is clearly directed from base to collocate. Nouns like *smoker* have a small number of typical collocates, whereas *heavy* does not select a particular noun.

Similar asymmetry effects are observed in human intuitions about associated words, which can be measured, e.g., with free association tasks (see Sec. 5.1). For instance, consider the pair (*mango*, *fruit*). When hearing the word *mango*, *fruit* is one of the first associations that come to mind. But when hearing *fruit*, more common fruits like *apple* are more likely to be the first associations rather than a less frequent fruit like *mango*. We call *fruit* → *mango* a *forward association* of the pair (*fruit*, *mango*) and a *backward association* of (*mango*, *fruit*). In this case, the forward association of (*fruit*, *mango*) is weak whereas its backward association is stronger.

We chose the terms *forward association* and *backward association* because we look at a broader class of associations between words than Sinclair, Kjellmer and Hausmann. While most collocations are lexically determined combinations of syntagmatically related words, human associations also include many paradigmatically related words (e.g. *boy* and *girl*). Note that statistical association measures have also been applied to the identification of such paradigmatic relations, in particular synonymy [19] and antonymy [6].

There are several reasons why human associations can be asymmetric. According to prototype theory [15], some members of a category are more prototypical than others. In our example, *apple* is a more prototypical example of fruit than *mango* (at least in North America), so that the directional association *fruit* → *apple* is stronger than *fruit* → *mango*.

Another possible reason for asymmetry is the degree of generality of terms. There is a tendency for a strong forward association from a specific term like *adenocarcinoma* to the more general term *cancer*, whereas the association from *cancer* to *adenocarcinoma* is weak.

We hypothesize that in many other cases the asymmetry is simply caused by frequency effects, corresponding to Sinclair's concepts of upward and downward collocation. For example, one of the most asymmetric pairs in our evaluation data is (*moo*, *cow*), with a very strong forward and a weak backward association. The word *moo* only occurs in the context of cows, but this is not true vice versa. Words like *milk* and *bull* are more frequent than *moo* in contexts where *cow* is used. This may not be an effect of prototypicality or specificity since the two terms are not related by a relationship such as hyponymy or meronymy.

Despite the strong intuitive support for the widespread existence of directed association provided by such examples, collocation studies still rely on symmetric association measures such as the well-known *pointwise MI*, *t-score* or *log-likelihood*. Our goal in this paper is to propose new measures that take asymmetric association into account and calculate separate scores for forward and backward association. We make use of psychological association norms (in particular, the *USF Free Association Database*, cf. Sec. 5.1) to evaluate how well these measures correspond to human intuitions about directed association.

It is important to distinguish "psychological association", the association of words in the human mind, which can be measured with reaction times and cue-target experiments, from "corpus association", the statistical association between terms in corpora. We will test in this paper to what extent directed corpus association can be used as a model for directed psychological association. We expect the results to carry over when the statistical measures are applied to collocation extraction tasks, for which no suitable (directed) reference data are currently available.

The paper is organized as follows. In Section 2, we discuss related work. Then, two asymmetric association measures are introduced in Section 3. Section 4 describes our methodology and corpus data. Section 5 presents results and evaluation, followed by a conlusion in Section 6.

2 Related work

In most cases, association measures are not used to examine the asymmetrical aspect of collocations. According to Evert [2, p. 75]:

> [t]he scores computed by an association measure can be interpreted in different ways: (i) They can be used directly to estimate the magnitude of the association between the components of a pair type.[1] (ii) They can be used to obtain a ranking of the pair types in the data set. In this case, the absolute magnitude of the score is irrelevant. (iii) They can also be used to rank pair types with a particular first or second component. [...]. I do not go further into (iii), which is closely tied to a "directional" view of cooccurrences and casts an entirely different light on the properties of association measures.

The first two approaches, which are symmetric, are predominant in computational linguistics [2] and statistical natural language processing [10]. In order to model the asymmetric relationship between two given words, this work focuses on the *directional* view which Evert [2, p. 27] describes as follows:

> An alternative is the "directional" view, which starts from a given *keyword* and aims to identify its *collocates*. [...] the evaluation of directional methods is more complicated and not as clear-cut. So far, published experiments have been limited to impressionistic case studies for a small number of keywords [1, 16, 18].

We are not aware of systematic research on asymmetrical association measures. In their comprehensive survey [13], Pecina and Schlesinger mention the two measures "conditional probability" and "reversed conditional probability", but do not discuss and evaluate them. Asymmetry has played a more important role in models of distributional similarity (which in turn have sometimes been used to model human associations), and several asymmetric similarity measures have been developed [4, 9, 11, 14]. Since these approaches focus on a different statistical aspect than association measures and cannot be compared directly, we do not go into further detail here.

3 The asymmetric association measures

3.1 Conditional probability

As our first measure, we use simple conditional probabilities, defined as the ratio between the joint probability of the pair and the probability of either the first word w_1 or the second word w_2:

$$P(w_2|w_1) = \frac{P(w_1, w_2)}{P(w_1)} \quad P(w_1|w_2) = \frac{P(w_1, w_2)}{P(w_2)} \qquad (1)$$

All probabilities are maximum-likelihood estimates without any smoothing. $P(w_2|w_1)$ is interpreted as a quantitative measure for the forward corpus assocation $w_1 \rightarrow w_2$ of the pair (w_1, w_2), and $P(w_1|w_2)$ as a measure for the backward association $w_2 \rightarrow w_1$.

Example: In the BNC, the conditional probabilities for the pair (*tomato*, *soup*) are: $P(\mathit{tomato}|\mathit{soup}) = 0.03194$ and $P(\mathit{soup}|\mathit{tomato}) = 0.05652$ (see Sec. 4 for details of our experimental setup). This conforms with the intuition that the forward association *tomato* → *soup* is stronger than the backward association *soup* → *tomato*.

3.2 Rank measure

We chose to base the rank measure on the χ^2 test because it is a well-established statistical test for association and is easy to implement. Using a different association measure would result in a different rank measure. To compute the rank measure, we first compute the X^2 statistic for each pair (w_1, w_2) in the corpus data as follows:

$$X^2(w_1, w_2) = \frac{O_{..} \cdot (O_{11}O_{22} - O_{12}O_{21})^2}{O_{1\cdot}O_{\cdot 1}O_{2\cdot}O_{\cdot 2}} \qquad (2)$$

[1] The term *pair type* refers to a representation of a collocation that is independent of surface form.

Using standard notation for contingency tables, O_{22} is the number of cooccurrence pair tokens that do not contain either of the two words, O_{12} the number containing only w_2, O_{21} the number with only w_1, and O_{11} the number with both words; $O_{1\cdot}$ is the number of tokens containing w_1 regardless of whether they also contain w_2, $O_{2\cdot}$ the number of tokens *not* containing w_1 regardless of w_2, etc.; and $O_{\cdot\cdot}$ is the total number of cooccurrence tokens.

For each w_1 a sorted association list is created that contains every pair $(w_1, \cdot)$ together with its association score X^2, sorted from highest to lowest association score. Then, the X^2 scores are replaced by ranks, i.e. natural numbers starting with 1. Figure 1 shows an example for the words *soup* and *tomato* that illustrates this procedure. The lists have been shortened to show only the relevant data.

If m consecutive w_2 have the same association score they are assigned the same rank r, and the w_2 with the next highest score is assigned rank $r + m$. We only consider the 1000 highest-ranked words in each list. We denote the rank of w_2 in the X^2 ranking of w_1 as follows:

$$R(w_2|w_1) \tag{3}$$

R is defined in analogy to conditional probability $P(w_2|w_1)$ which returns the probability of seeing w_2 when w_1 has already appeared. Analogously, $R(w_2|w_1)$ returns the rank of w_2 in the association list of w_1. Using the information in Figure 1, the ranks for the example pair (*soup*, *tomato*) can be determined. They are $R(\textit{tomato}|\textit{soup})$ ("tomato given soup") $= 3$ and $R(\textit{soup}|\textit{tomato})$ ("soup given tomato") $= 10$. A lower rank indicates stronger association, hence the rank measure shows a stronger association for *soup* → *tomato* than for *tomato* → *soup*.

We note that the asymmetric rank measure is based on a symmetric association measure, the χ^2 test. According to the χ^2 test, the pairs *(mango, fruit)* and *(fruit, mango)* have the same association strength because the measure is not directed. But *fruit* will figure more prominently in the association list of *mango* ($R(\textit{fruit}|\textit{mango}) = 10$) than vice versa ($R(\textit{mango}|\textit{fruit}) = 47$). The rank measure proposed here uses this type of difference in the associational rankings to transform symmetric χ^2-based association scores into asymmetric R measure ranks.

4 Methodology

We selected the *British National Corpus* (BNC) (http://www.natcorp.ox.ac.uk/) as a data set for calculating corpus associations. It is a large balanced corpus of approximately 100 million words, containing samples from various genres and sources such as newspapers, popular fiction and scientific journals. Words and punctuation tokens have been automatically annotated with syntactic categories, using the *BNC Basic Tagset*. These annotations make it easy to apply a part-of-speech filter to the cooccurrence data.

In order to extract data that provide information about the corpus association between words, cooccurrence pairs were constructed in the following way: First, words containing special characters (e.g., é, £ or "/") and words starting with numbers or other non-letter characters were excluded from the experiment. In addition, words shorter than three characters were ignored. Each word in the corpus was then combined with its ten predecessors as well as its ten successors. A part-of-speech filter was applied, allowing only adjectives, nouns and proper nouns in the pairs. No further linguistic processing was done except for lowercasing of sentence-initial words that were not tagged as proper nouns. Subsequently, all words that occur less than 40 times in the BNC were discarded (together with the corresponding pairs). In this way, a list of 28,149,644 word tokens and 177,913,470 cooccurrence pairs (i.e., not necessarily distinct tokens of word pairs) was obtained.

5 Results and evaluation

The asymmetric measures were evaluated using two different methods. First, the forward and backward associations calculated for highly asymmetric and symmetric pairs were calculated. For this purpose, the ten most asymmetric and the ten most symmetric pairs were extracted from a reference set of human "psychological" association (see Sec. 5.1). Second, the ability of the measures to predict the asymmetry or symmetry of given pairs was evaluated on a large set of 5697 word pairs from the reference database.

5.1 Reference data

The performance of the two asymmetric association measures is evaluated against word pairs from a database that contains the results of free association experiments. This database, the *University of South Florida Free Association Norms* [12], consists of labeled *cue-target pairs* where a *cue* is a word presented to a subject and the corresponding *target* is the word that the subject wrote down on a blank shown next to the cue. The experiment is described as follows:

> Participants were asked to write the first word that came to mind that was meaningfully related or strongly associated to the presented word on the blank shown next to each item. [...] For example, if given BOOK ________, they might write READ on the blank next to it. This procedure is called a discrete association task because each participant is asked to produce only a single associate to each word.

Each of the 5,019 cue words is listed in the database together with all the targets that subjects produced for it. For every single cue-target pair, a database entry lists how many subjects were presented the cue and how many of them named each target. Figure 2 shows two abbreviated entries. The number of test persons that were presented a cue word is labeled #G. #P is the number of people that gave a particular target response. The *forward strength* (FSG) is #P divided by #G and the *backward strength* (BSG) is the forward strength of the reversed pair. The terms FSG and BSG were introduced by the creators of the *USF Free Association Norms.*

Our evaluation methodology is most similar to that of [8, 19] who evaluate corpus-derived association measures on a synonym gold standard that reflects human

w_1	X^2 score	w_2	$R(w_2\|w_1)$
soup	14666.277	bowl	1
	14531.099	pea	2
	7681.563	*tomato*	3
	6082.888	kitchens	4
	4116.237	mushroom	5

w_1	X^2 score	w_2	$R(w_2\|w_1)$
tomato	8770.224	pepper	8
	8531.046	cucumber	9
	7681.563	*soup*	10
	7594.471	salad	11
	7417.416	chopped	12

Fig. 1: *Ranking applied to X^2 scores of the words* soup *and* tomato

CUE	TARGET	#G	#P	FSG	BSG
aardvark	anteater	152	9	.059	.117
anteater	aardvark	145	17	.117	.059

Fig. 2: *Example from the* Free Association Norms

understanding of synonymy. However, their gold standard has a single correct answer (out of a small number of alternatives) for each cue word, rather than a large number of target words with different degrees of (forward) association.

5.2 Strong asymmetric associations

The first part of the evaluation is concerned with analyzing the performance of the asymmetric measures for pairs that are highly asymmetric in the reference data set. In order to create a suitable reference list from the association database, cue-target pairs with a high difference between FSG and BSG were extracted. The absolute value of the difference had to be greater than 0.7 for a pair to be selected. In order to make the list comparable to the results that are based on the corpus data, the list had to be filtered: First, all cue-target pairs with BSG 0 were removed. In those cases, the test persons were never presented the target word as a cue word so there is not enough data to determine both FSG *and* BSG. The second step eliminated parts of speech that were not included during the processing of the corpus data. All pairs containing words that do not occur in the corpus (or did not pass the filters) were eliminated as well.

Figure 3 shows the ten most asymmetric cue-target pairs together with the ranks and conditional probabilities that were computed from the BNC data. Obviously, conditional probabilities correspond much better to the human ratings than the rank measure. FSG exceeds BSG roughly by a factor of 10 for all ten pairs, and the conditional probabilities mirror this relation. In eight out of ten pairs that were evaluated, the ratio between $P(w_2|w_1)$ and $P(w_1|w_2)$ is on the same order of magnitude as the ratio between FSG and BSG. In two cases, namely pairs 4 and 8, the results deviate slightly from this pattern. The latter shows a ratio of about 4, the former a ratio of approximately 174.

The comparison between FSG, BSG, and the rank measure is less straightforward. Two quantities have to be taken into account: The difference between $R(w_2|w_1)$ and $R(w_1|w_2)$, as well as the absolute value of $R(w_2|w_1)$. First, in order for the rank measure to express that forward association is stronger than backward association, $R(w_2|w_1)$ must be lower than $R(w_1|w_2)$. Second, $R(w_2|w_1)$ should be small in order to express *strong* forward association. However, only seven out of the ten pairs satisfy the first condition, and only five of them also meet the second criterion. The other two (number 4 and number 7) have forward ranks of 35 and 47, respectively, which do not indicate strong forward association. Two pairs (numbers 1 and 8) are almost symmetric according to the rank measure and pair number 3 even shows a weak backward association.

Another difficulty with the rank measure is the interpretation of the magnitude of the *rank difference* $\delta = |R(w_2|w_1) - R(w_1|w_2)|$. First, it is not clear how large the value of δ has to be in order to indicate strong asymmetry and second, it can only be interpreted in combination with the absolute ranks. E.g., word pairs 1 and 9 both have a rank difference of 2, but this difference is arguably more "important" for pair 9 (rank 2 vs. rank 4) than for pair 1 (rank 7 vs. rank 9).

Conditional probabilities correctly predict the direction of the asymmetry in all 10 cases. The rank measure only predicts the correct direction in 7 out of 10 cases.

5.3 Strong symmetric association

In addition to the strongly asymmetric pairs discussed in the last section, the reference database contains pairs with symmetric associations, i.e., FSG and BSG are almost equal. Although the measures presented in this work aim at capturing the asymmetry in the human associations, they should also be able to predict word pairs with *symmetric* associations and distinguish them from the asymmetric ones.

Symmetric pairs were extracted from the reference data by selecting pairs with $|\text{FSG} - \text{BSG}| < 0.1$ and $\text{FSG} > 0.5$ (in order to remove weakly associated pairs). Again, the list was filtered based on part of speech and occurrence of the words in the BNC data (cf. 5.2). Then the ten most symmetric pairs (i.e. those with the smallest difference between FSG and BSG) were evaluated. The results are shown in Fig. 4.

Symmetry is reflected by the two measures in an entirely different manner than asymmetry. The conditional probabilities did not match FSG/BSG ratios as well as in Section 5.2, while the rank measure achieved slightly better results for strongly symmetric pairs than for asymmetric pairs.

In order for conditional probabilities to express strong symmetric association, their quotient should be close to 1. However, the only word pairs meeting this requirement to some extent are pairs 1 and 4. In all other cases there is a strong discrepancy between

[2] The American English *omelet* appears as *omelette* in the BNC.

No.	w_1	w_2	FSG − BSG	FSG	BSG	$R(w_2\|w_1)$	$R(w_1\|w_2)$	$P(w_2\|w_1)$	$P(w_1\|w_2)$	$\approx \frac{P(w_2\|w_1)}{P(w_1\|w_2)}$
1	trout	fish	0.877	0.913	0.036	9	7	0.15987	0.01042	15
2	Cheddar	cheese	0.867	0.922	0.055	2	7	0.29906	0.01331	22
3	exhausted	tired	0.82	0.895	0.075	104	87	0.01479	0.00139	10
4	crib	baby	0.81	0.842	0.032	35	69	0.10638	0.00061	174
5	omelet[2]	eggs	0.809	0.836	0.027	3	26	0.16513	0.00504	32
6	wick	candle	0.79	0.841	0.051	3	5	0.08823	0.00807	11
7	teller	bank	0.786	0.814	0.028	47	84	0.07438	0.00099	75
8	bank	money	0.78	0.799	0.019	11	10	0.05767	0.01449	4
9	saddle	horse	0.776	0.879	0.103	2	4	0.11467	0.00997	11
10	bouquet	flowers	0.775	0.828	0.053	1	4	0.21862	0.01108	19

Fig. 3: *Comparison of strong asymmetric human association with associations computed from corpus data*

No.	w_1	w_2	|FSG − BSG|	FSG	BSG	$R(w_2\|w_1)$	$R(w_1\|w_2)$	$P(w_2\|w_1)$	$P(w_1\|w_2)$	$\approx \frac{P(w_2\|w_1)}{P(w_1\|w_2)}$
1	boys	girls	0.003	0.500	0.503	1	1	0.17965	0.14873	1.20
2	happy	sad	0.006	0.628	0.634	9	4	0.00725	0.02412	0.30
3	pepper	salt	0.006	0.695	0.701	1	1	0.48230	0.13897	3.47
4	legs	arms	0.008	0.541	0.549	2	1	0.07842	0.04870	1.61
5	bad	good	0.008	0.750	0.758	4	2	0.11129	0.02083	5.34
6	dinner	supper	0.01	0.535	0.545	45	16	0.00455	0.02037	0.22
7	grandma	grandpa	0.015	0.538	0.553	2	3	0.03333	0.1375	0.24
8	negative	positive	0.024	0.603	0.627	1	1	0.20472	0.10928	1.87
9	closing	opening	0.027	0.480	0.507	19	9	0.02495	0.00445	5.60
10	far	near	0.032	0.503	0.535	10	6	0.00898	0.02282	0.39

Fig. 4: *Comparison of strong symmetric human association with associations computed from corpus data*

$P(w_1|w_2)$ and $P(w_2|w_1)$, so that the conditional probabilities fail to capture the high symmetry that the reference data suggest.

The performance of the rank measure was better in that it accurately predicted the symmetry of the pairs in six cases (pairs 1, 3, 4, 5, 7 and 8). In three cases, the ranks accurately indicated both perfect symmetry ($R(w_2|w_1) = R(w_1|w_2) = 1$) and a strong association between the two words (because of the low rank). For pairs 2, 9 and 10, high ranks $R(w_2|w_1)$ indicate that there is no strong forward association. While $R(w_1|w_2)$ is lower in each case, the backward associations are not strong enough to conclude that the pairs are clearly identified as asymmetric. In particular, pair 10 has quite similar forward and backward ranks and may be considered near-symmetric. For pair 6, the rank measure indicates a clearly asymmetric, but overall weak association.

The rank measure predicts symmetry or near-symmetry (defined as a rank difference $\delta \leq 5$) for 8 of the 10 test pairs. Conditional probabilities perform less well and only predict symmetry or near-symmetry (defined as $0.5 \leq \frac{P(w_2|w_1)}{P(w_1|w_2)} \leq 2$) for 3 out of 10 pairs.

5.4 Automated evaluation

To evaluate the two asymmetric measures on a larger scale, we extracted all pairs (w_1, w_2) that occur in both directions in the USF data set. We then selected the direction with FSG>BSG. The resulting set was randomly split into a training set consisting of 3000 pairs and a test set consisting of 2697 pairs. We then determined the median of the FSG−BSG values (0.049) and evaluated the asymmetric measures on their ability to predict whether FSG−BSG was ≥ 0.049 (intuitively understood as asymmetric pairs) or < 0.049 (understood as symmetric pairs).

We used logistic regression in R[3] for predictive analysis. The response variable is FSG−BSG $\geq 0.049/< 0.049$. Initially, we intended to use either the two ranks or the two conditional probabilities as predictive variables. In preliminary experiments on the training data we found that a log transformation of the predictor variables improved the model. We therefore used the logs of ranks / conditional probabilities as predictors instead of the original variables.

When applied to the test set, accuracies of predicting symmetry (FSG−BSG < 0.049) vs. asymmetry (FSG−BSG ≥ 0.049) were 59% for ranks, 61% for conditional probabilities and 62% for a combination of ranks and conditional probabilities. All three results are significantly different from the baseline accuracy of 50% ($p < 0.001$, χ^2 test). The three results were not significantly different from each other (e.g., $p = 0.4243$ for ranks vs. conditional probabilities, χ^2 test).

We conclude the following from this evaluation: (i) Both measures contain information about "psychological" asymmetry. (ii) There is no significant difference in accuracy of prediction between the two measures. (iii) Overall accuracy is low. This is partly due to the general difficulty of modeling human judgments with corpus data, but it may also indicate that there are more effective measures of asymmetry than the ones we have investigated here. The data sets are available at `http://ifnlp.org/ranlp07`.

6 Conclusion and future work

We introduced two asymmetric statistical association measures that aim to capture the asymmetry of human

[3] `http://www.r-project.org/`

word associations, one based on conditional probabilities and the other on ranks according to an established association measure. Both measures were implemented and applied to a large data set of cooccurrences extracted from the *British National Corpus*. The resulting directed association scores were evaluated against norms obtained from free association tasks with human subjects (the *USF Free Association Norms* database).

We found that the new measures are able to distinguish between symmetric and asymmetric word pairs to some extent, but with a relatively high error rate (62% accuracy vs. 50% baseline). Additional experiments with a small number of highly symmetric and highly asymmetric pairs showed that the measure based on conditional probabilities works well for asymmetric pairs and makes reasonable predictions for the magnitude of the asymmetry. However, its scores for highly symmetric pairs were unreliable and difficult to interpret. The rank-based measure seems more suitable for identifying symmetric pairs. It is also the more robust measure overall, with $\geq 50\%$ accuracy for both sets.

The work presented here can be extended in many ways. Our evaluation results are encouraging, but show that there is considerable room for improvement. Some extensions are concerned with the definitions of the asymmetric association measures. The maximum-likelihood estimates used by the conditional probability measure could be replaced by smoothed estimates or confidence intervals. Then rank-based measure, which currently uses rankings according to the X^2 statistic, can equally well be based on any other standard association measure. Further research is also needed on the interpretation of rank differences.

Performance of the measures might be improved by working on lemmatized data, which is offered by the new XML edition of the BNC. This would help to abstract over surface forms, thus "tidying up" the association lists and increasing the significance of statistical association. Experiments with different window sizes and filtering constraints can also be performed. Finally, scaling up to much larger Web corpora would further boost statistical significance and produce more reliable association scores.

References

[1] K. W. Church and W. A. Gale. Concordances for parallel text. In *Proceedings of the 7th Annual Conference of the UW Center for the Nwe OED and Text Research*, 1991. Quoted in Evert.

[2] S. Evert. *The Statistics of Word Cooccurrences - Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart, 2004.

[3] J. R. Firth. A Synopsis of Linguistic Theory, 1933-1955. In J. R. Firth, editor, *Studies in Linguistic Analysis*. Blackwell, Oxford, 1957.

[4] M. Geffet and I. Dagan. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the ACL 2006*. Association for Computational Linguistics, 2005.

[5] F. J. Hausmann. Le dictionnaire de collocations. In *In Wrterbcher, Dictionaries, Dictionnaires. Ein internationales Handbuch*. de Gruyter, Berlin, 1989.

[6] J. S. Justeson and S. M. Katz. Co-occurrences of antonymous adjectives and their contexts. *Computational Linguistics*, 17(1), 1991.

[7] G. Kjellmer. A mint of phrases. In K. Aijmer and B. Altenberg, editors, *English Corpus Linguistics*. Longman, London, 1991.

[8] T. K. Landauer and S. T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 1997.

[9] L. Lee. Measures of Distributional Similarity. In *37th Annual Meeting of the Association for Computational Linguistics*, 1999.

[10] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 1999.

[11] D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. Finding predominant word senses in untagged text. In *Proceedings of the ACL 2004*. Association for Computational Linguistics, 2004.

[12] D. L. Nelson, C. L. McEvoy, and T. A. Schreiber. The University of South Florida word association, rhyme, and word fragment norms. http://www.usf.edu/FreeAssociation/, 1998.

[13] P. Pecina and P. Schlesinger. Combining association measures for collocation extraction. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 651–658, 2006.

[14] V. Pekar. Acquisition of verb entailment from text. In *Proceedings of the Human Language Technology/North American Association for Computational Linguistics (HLT/NAACL-06)*, 2006.

[15] E. H. Rosch. Natural Categories. *Cognitive Psychology*, 4, 1973.

[16] J. Sinclair. *Corpus, Concordance, Collocation*. Oxford Universtity Press, Oxford, 1991. Quoted in Evert.

[17] J. Sinclair, S. Jones, R. Daley, and R. Krishnamurthy. *English Collocation Studies: The OSTI Report*. Continuum Books, London and New York, 2004.

[18] M. Stubbs. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, 2(1), 1995. Quoted in Evert.

[19] P. D. Turney. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167, 2001.