A Comparative Evaluation of Collocation Extraction Techniques

Darren Pearce

School of Cognitive and Computing Sciences (COGS)
University of Sussex
Falmer
Brighton
BN1 9QH
darrenp@cogs.susx.ac.uk

Abstract

This paper describes an experiment that attempts to compare a range of existing collocation extraction techniques as well as the implementation of a new technique based on tests for lexical substitutability. After a description of the experiment details, the techniques are discussed with particular emphasis on any adaptations that are required in order to evaluate it in the way proposed. This is followed by a discussion on the relative strengths and weaknesses of the techniques with reference to the results obtained. Since there is no general agreement on the exact nature of collocation, evaluating techniques with reference to any *single* standard is somewhat controversial. Departing from this point, part of the concluding discussion includes initial proposals for a common framework for evaluation of collocation extraction techniques.

1. Introduction

Over the last thirty years, there has been much discussion in the linguistics literature on the exact nature of collocation. Even now, there is no widely-accepted definition.

A by-product of the lack of any agreed definition is a lack of any consistent evaluation methodology. For example, Smadja (1993) employs the skills of a professional lexicographer, Blaheta and Johnson (2001) use native speaker judgements and Lin (1998) uses a term bank. Evaluation can also take the form of a discussion of the 'quality' of the extracted collocations (e.g. Kita and Ogata (1997)).

Since it is somewhat controversial to evaluate these techniques with reference to anything considered as a standard whether relying on human judgement or machine readable resources, relatively little work has been done on comparative evaluation in this area (e.g. Evert and Krenn (2001), Krenn and Evert (2001) and Kita et al. (1994)). However, the view taken in this paper is that such evaluation does offer at least *some* departure points for a discussion of the relative merits of each technique.

2. Existing Collocation Extraction Techniques

Various researchers in natural language processing have proposed computationally tractable definitions of collocation accompanying empirical experiments seeking to validate their formulation. With the availability of large corpora came several techniques based on N-gram statistics derived from these large bodies of text. This begins with the z-score of Berry-Rogghe (1973) and later with Church and Hanks (1990), Kita et al. (1994) and Shimohata et al. (1997).

Smadja (1993) went one step further and developed a technique that not only extracted continuous sequences of words that form collocations but also those with gaps (e.g. $\langle$*break-down*, *door*$\rangle$ in *break down the old door*). Underpinning the technique was the implicit inference of syntax through analysis of co-occurrence frequency distributions.

With recent significant increases in parsing efficiency and accuracy, there is no reason why explicit parse information should not be used. Indeed, Goldman et al. (2001) emphasises the necessity to use such information since some collocational dependencies can span as many as 30 words (in French). Several researchers have exploited the availability of parse data. Lin (1998) uses an information-theoretic approach over parse dependency triples, Blaheta and Johnson (2001) use log-linear models to measure association between verb-particle pairs and Pearce (2001) uses restrictions on synonym substitutions within parse dependency pairs.

Recent research into collocation extraction (Pearce, 2001) has produced a new technique based on analysing the possible substitutions for synonyms within candidate phrases. For example, *emotional baggage* (a collocation) occurs more frequently than the phrase *emotional luggage* formed when *baggage* is substituted for its synonym *luggage*. This technique uses WordNet (Miller, 1990) as a source of synonymic information.

3. Experiment Description

The experiment compares implementations of many of the techniques discussed in the previous section. Each technique was supplied with bigram co-occurrence frequencies obtained from 5.3 million words of the BNC. For each pair of words in this frequency data, the technique was used to assign a 'score' where the higher the score, the 'better' the collocation (according to the technique). Alternatively, the technique, based on some predefined criteria (such as the application of a threshold), could opt to ignore scoring the pair. Such a strategy tends to increase precision at the expense of recall.

Evaluation of the output list uses the multi-word information in the machine-readable version of the New Oxford Dictionary of English (NODE) (Pearsall and Hanks, 1998). This was processed to produce a list of 17,485 two-word collocations such as *adhesive tape*, *hand grenade* and *value judgement*. This list was subsequently reduced to 4,152 en-

tries that occurred at least once in the same data supplied to the techniques, thus forming the 'gold' standard.

4. Technique Details

The following subsections briefly describe each of the evaluated techniques with particular emphasis on the assignment of scores to word sequences, especially sequences of length two. In addition, accompanying each description is a figure consisting of a small example of the application of the technique to a fabricated corpus of a 1000 words. The subsection titles also serve to identify the way in which the techniques will be referred to in Section 5.

Throughout the descriptions, a word, w (x, y, etc), when part of a (contiguous) sequence of words, $\mathbf{w} = \langle w_1 \dots w_n \rangle$ (or $w_{1,n}$), may be written w_i indicating its position within this sequence where $1 \leq i \leq n$ and $|\mathbf{w}| = n$. It is useful to distinguish the frequency of a word, $f(w)$, the frequency of a sequence, $f(\mathbf{w})$, and, in particular, the distance frequency, $f^d(x, y)$, which represents the count of word y occurring d words after x. In general. the distance, d, can be negative as well as positive, represented by the set $\mathcal{D} = \{-n, \dots 1, 1, \dots n\}$. When this set is restricted to a positive range only, this is written $\mathcal{D}^+ = \{1 \dots n\}$. The arithmetic mean of a bag, $\mathcal{A}$ is notated $\mu(\mathcal{A})$ and the standard deviation by $\sigma(\mathcal{A})$.[1]

4.1. Berry-Rogghe (1973) (berry)

One of the earliest attempts at automatic extraction, this technique uses a window of words. Given a word, x, and its frequency, $f(x)$, the probability of another word, y, occurring in any other position in the corpus is approximated by:

$$p(y) = \frac{f(y)}{N - f(x)}$$

where the normalisation factor ($N - f(x)$) is the number of words in the corpus that aren't x. The *expected* frequency of x and y within a certain window of words is then:

$$\hat{f}(x, y) = p(y) \cdot f(x) \cdot |\mathcal{D}|$$

in which there are $|\mathcal{D}|$ chances around each x for y to occur. The significance of the *actual* frequency count, $f(x, y)$ in comparison to $\hat{f}(x, y)$ is computed using a normal approximation to the binomial distribution, yielding the z-score:

$$z = \frac{f(x, y) - \hat{f}(x, y)}{\sqrt{\hat{f}(x, y) \cdot (1 - p(y))}}$$

In order to process bigram data, this must be modified slightly such that the window is just one word wide.

[1]These functions, defi ned purely for the sake of brevity, necessitate the use of bag-theory. Bags are similar to sets except that they can contain repeated elements. To correspond with this similarity, they are written using the same font as for sets ($\mathcal{A}$, $\mathcal{B}$, etc). The distinction is made where appropriate although this is usually obvious from the context.

```
  ... open the door ... but the door
was open ... so I left the door open ...
open ... the open door ...
```

$f(\textit{open}) = 5$ and $f(\textit{door}) = 4$ so:

$$p(\textit{door}) = \frac{f(\textit{door})}{N - f(\textit{open})} = \frac{4}{1000 - 5} = 0.004$$

and

$$\begin{aligned}\hat{f}(x, y) &= p(\textit{door}) \cdot f(\textit{open}) \cdot |\mathcal{D}| \\ &= 0.004 \times 5 \times 4 = 0.08\end{aligned}$$

With $f(x, y) = 4$:

$$z = \frac{4 - 0.08}{\sqrt{0.08(1 - 0.004)}} = 13.9$$

Figure 1: Example for Berry-Rogghe (1973).

4.2. Church and Hanks (1990) (church)

Although widely used as a basis for collocation extraction, Church and Hanks (1990) in fact measured word *association*. The probability of seeing words x and y within a pre-defined window is given by:

$$p(x, y) = \frac{\sum_{d \in \mathcal{D}^+} f^d(x, y)}{|\mathcal{D}^+| \cdot N}$$

where the normalisation factor takes account of the possibility of counting the same word more than once within the window.[2] The co-occurrence of the two words is scored using point-wise mutual information:

$$I(x, y) = \log_2 \frac{p(x, y)}{p(x) \cdot p(y)}$$

where word probabilities are calculated directly using relative frequency. This gives the amount of information (in bits) that the co-occurrence gives over and above the information of the individual occurrences of the two words.

Since this measure becomes unstable when the counts are small, it is only calculated if $f(x, y) > 5$.

4.3. Kita et al. (1994) (kita)

This technique is based on the idea of the cognitive cost of processing a sequence of words, $c(\mathbf{w})$. In general, the non-collocational cost of a word sequence is assumed to be linear in the number of words ($c_w(\mathbf{w}) = |\mathbf{w}|$).[3] However, collocations are processed as a single unit ($c_c(\mathbf{w}) = 1$). Motivated by the rationale that a collocation would serve to *reduce* the cognitive cost of processing a word sequence, collocations are extracted by considering cost reduction

[2]Note that $|\mathcal{D}^+|$ corresponds to $w - 1$ in Church and Hanks (1990).

[3]This is, as the authors admit, a greatly simplifying assumption.

```
... butterfly stroke ... butterfly ...
stroke ... butterfly stroke ... stroke
... stroke ... stroke ...
```

$$p(\textit{butterfly}) = \frac{3}{1000} \qquad p(\textit{stroke}) = \frac{6}{1000}$$

so, using a window of two words:

$$p(\textit{butterfly stroke}) = \frac{2}{2 \times 1000}$$

and

$$I(\textit{butterfly stroke}) = \log_2 \frac{0.001}{0.003 \times 0.006} = 5.8$$

Figure 2: Example for Church and Hanks (1990).

```
... animal liberation ... animal lib-
eration front ... animal liberation ...
animal liberation ... animal liberation
front ...
```

The table below shows the cost reduction calculations for two word sequences: *animal liberation* ($\mathbf{w}_1$) and *animal liberation front* ($\mathbf{w}_2$).

$\mathbf{w}$	$\lvert\mathbf{w}\rvert$	f	C_w	C_c	K
$\mathbf{w}_1$	2	5	10	5	5
$\mathbf{w}_2$	3	2	6	2	4

However, since $\mathbf{w}_1 \sqsubset \mathbf{w}_2$, the calculations above have considered two occurrences of $\mathbf{w}_2$ also as occurrences of a larger collocation, $\mathbf{w}_1$. The cost reduction of $\mathbf{w}_1$ must therefore be reduced accordingly:

$$K'(\mathbf{w}_1) = 1 \cdot (f(\mathbf{w}_1) - f(\mathbf{w}_2)) = 5 - 2 = 3$$

Figure 3: Example for Kita et al. (1994).

across all occurrences of the phrase within the corpus. With the two costs calculated by:

$$C_w(\mathbf{w}) = c_w(\mathbf{w}) \cdot f(\mathbf{w}) = |\mathbf{w}| \cdot f(\mathbf{w})$$
$$C_c(\mathbf{w}) = c_c(\mathbf{w}) \cdot f(\mathbf{w}) = f(\mathbf{w})$$

the difference represents the cost reduction, $K(\mathbf{w})$, for processing the word sequence as a unit rather than as separate words:

$$\begin{aligned} K(\mathbf{w}) &= C_w(\mathbf{w}) - C_c(\mathbf{w}) \\ &= (|\mathbf{w}| - 1) \cdot f(\mathbf{w}) \end{aligned}$$

The set of collocations, $\mathcal{C}$, extracted by this method is formed by applying a threshold to the cost reduction:

$$\mathcal{C} = \{\mathbf{w} : \ K(\mathbf{w}) > T\}$$

Special consideration must be made, however, for the possibility of one collocational phrase as a subsequence of another. This leads to the corrected cost reduction, K':

$$K'(\mathbf{w}) = (|\mathbf{w}| - 1) \cdot (f(\mathbf{w}) - \sum_{\substack{\mathbf{x} \in \mathcal{C}, \\ \mathbf{w} \sqsubset \mathbf{x}}} f(\mathbf{x}))$$

where $\sqsubset$ is the subsequence operator. Importantly, when $|\mathbf{w}| = 2$, the cost reduction is just the frequency of the word sequence: $K(\mathbf{w}) = K'(\mathbf{w}) = f(\mathbf{w})$.

4.4. Shimohata et al. (1997) (shim)

By determining the entropy of the immediate context of a word sequence, this technique ranks collocations according to the assumption that collocations occur as units in (an information-theoretically) 'noisy' environment.[4] The probability distributions of the words on each side:

$$p(w, \mathbf{w}) = \frac{f(w, \mathbf{w})}{f(\mathbf{w})} \qquad p(\mathbf{w}, w) = \frac{f(\mathbf{w}, w)}{f(\mathbf{w})}$$

are used to find the left and right entropy:

$$H(\bullet\mathbf{w}) = \sum_{i=1}^{n_l} -p(w_i, \mathbf{w}) \log_2 p(w_i, \mathbf{w})$$
$$H(\mathbf{w}\bullet) = \sum_{j=1}^{n_r} -p(\mathbf{w}, w_j) \log_2 p(\mathbf{w}, w_j)$$

where n_l and n_r are the number of different words occurring to the left and the right of $\mathbf{w}$ respectively. The entropy of the word sequence is then given by:

$$H(\mathbf{w}) = min(H(\bullet\mathbf{w}), H(\mathbf{w}\bullet))$$

4.5. Blaheta and Johnson (2001) (blaheta)

In the specific case of two-word collocations, this technique corresponds to the log odds ratio. Performed on a 2×2 contingency table, this measure indicates the degree to which one variable influences another and is unaffected by sample size (Howell, 1992). For a particular pair of words, $\langle w_1, w_2 \rangle$, the contingency table, based on pair counts of the form $\langle x, y \rangle$ is then:

	$y \neq w_2$	$y = w_2$	
$x \neq w_1$	c_0	c_2	$c_0 + c_2$
$x = w_1$	c_1	c_3	$c_1 + c_3$
	$c_0 + c_1$	$c_2 + c_3$	$c_0 + c_1 + c_2 + c_3$

where the counts $c_{0,\ldots,3}$ can be calculated directly from frequency information.[5] The odds of w_2 occurring given w_1

[4] In fact, this forms only the first phase of the technique proposed in Shimohata et al. (1997). The second phase combines these collocational units iteratively (after applying a threshold) to yield collocations with 'gaps'. The technique is therefore capable of extracting both interrupted and uninterrupted collocations. For the purposes of the task investigated in this paper, the second phase is ignored.

[5] In practice, the frequencies are corrected for continuity by adding 0.5.

```
... actual bodily harm ... grievous
bodily harm ... actual bodily harm ...
grievous bodily harm ... grievous bod-
ily harm ... grievous bodily harm ...
grievous bodily harm ... actual bodily
harm
```

$$f(\textit{bodily harm}) = 8$$
$$f(\textit{actual bodily harm}) = 3$$
$$f(\textit{grievous bodily harm}) = 5$$

so

$$p(\textit{actual}, \textit{bodily harm}) = \frac{3}{8}$$
$$p(\textit{grievous}, \textit{bodily harm}) = \frac{5}{8}$$

giving:

$$\begin{aligned} H(\bullet\mathbf{w}) &= \sum_{i=1}^{n_l} -p(w_i, \mathbf{w}) \log_2 p(w_i, \mathbf{w}) \\ &= -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} \\ &= 0.95 \end{aligned}$$

Figure 4: Example for Shimohata et al. (1997).

has *not* occurred is c_2/c_0. Similarly, the odds of w_2 occurring given w_1 *has* occurred is c_3/c_1. The ratio of this second fraction to the first is the odds ratio:

$$\frac{c_3/c_1}{c_2/c_0} = \frac{c_0 c_3}{c_1 c_2}$$

Taking the natural logarithm of this measure gives the log odds ratio:

$$\lambda = \log\frac{c_0 c_3}{c_1 c_2} = \log c_0 - \log c_1 - \log c_2 + \log c_3$$

which has a standard error of:

$$\sigma = \sqrt{\sum_{b=0}^{3}\frac{1}{c_b}}$$

Collocations are then scored by calculating $\mu = \lambda - 3.29\sigma$ which sets μ to the lower bound of a 0.001 confidence interval. This has the effect of trading recall for precision.

4.6. Pearce (2001) (pearce)

This supervised technique is based on the assumption that if a phrase is semantically compositional, it is possible to substitute component words within it for synonyms without a change in meaning. If a phrase does not permit such substitutions then it is a collocation. More specifically to

Assuming the 1000 word corpus consists of 40 sentences each 25 words in length, the total number of bigrams, $\sum_{i=0}^{3} c_i = 24 \times 40 = 960$. Given the following contingency table for the bigram collocation, *paddling pool*:

	$y \neq pool$	$y = pool$	
$x \neq paddling$	900	30	930
$x = paddling$	20	10	30
	920	40	960

the odds ratio is:

$$\frac{900 \times 10}{20 \times 30} = 15$$

and so $\lambda = \log 15 = 2.71$. The standard error in this estimate is:

$$\sqrt{\frac{1}{900} + \frac{1}{20} + \frac{1}{30} + \frac{1}{10}} = 0.43$$

so $\mu = \lambda - 3.29\sigma = 1.30$.

Figure 5: Example for Blaheta and Johnson (2001).

the experiment described in this paper, the *degree* to which such changes are possible is obtained and this is used as a score.

Mapping a word to its synonyms for each of its senses is achieved using WordNet, $\mathcal{W}$, which consists of a set of *syn*onym *sets* (or synsets):

$$\mathcal{W} = \{\mathcal{S}_1, \mathcal{S}_2, \dots\}$$

.

A phrase, $P = \langle w_1 \dots w_n \rangle$ is considered to be one lexical realisation of the concept P. Assuming that P is semantically compositional, the set of competing lexical realisations of P, $\mathcal{R}(P)$ can be constructed through reference to WordNet synsets:

$$\mathcal{R}(P) = \left\{w'_{1,n} : \ \forall i, 1 \leq i \leq n.\ w'_i \in \mathcal{S}_i\right\}$$

where the WordNet synset corresponding to the correct sense of each w_i is $\mathcal{S}_i$.

For each phrase $P_i \in \mathcal{R}(P)$, two probabilities are calculated. The first is the likelihood that the phrase will occur based on the joint probability of independent trials from each synset:

$$\hat{p}(P_i) = \prod_{i=1}^{n} p_i(w_i)$$

with $p_i(w_i)$, the probability that w_i is 'picked' from $\mathcal{S}_i$, approximated by a maximum likelihood estimate:

$$p_i(w_i) = \frac{freq(w_i|\mathcal{S}_i)}{\sum_{w \in \mathcal{S}_i} freq(w|\mathcal{S}_i)}$$

The second probability is based on the occurrences of

phrases (rather than words):

$$p(P_i) = \frac{f(P)}{\sum_{P' \in \mathcal{R}(P)} f(P')}$$

The difference between these two probabilities,

$$d_i = p(P_i) - \hat{p}(P_i)$$

indicates the amount of 'deviation' from what is expected under the substitution assumptions of this technique and what actually occurs. To gauge the strength of this deviation (and hence the strength of the collocation), each d_i is converted into units of standard deviation:[6]

$$z_i = \frac{d_i}{\sigma(d)}$$

In stark contrast to the other techniques investigated in this paper, this development of the work in Pearce (2001) requires sense information. In the absence of such data, the realisation function is modified to:

$$\widetilde{\mathcal{R}}(P) = \left\{ w'_{1,n} : \forall i, 1 \leq i \leq n.\ w'_i \in \mathcal{S} \ \wedge \ \mathcal{S} \in \mathcal{C}(w_i) \right\}$$

where $\mathcal{C}(w)$ is the concept set of w:

$$\mathcal{C}(w) = \{ \mathcal{S} : \ w \in \mathcal{S} \}$$

Similar modifications are also made to the calculation of $p(w_i)$. In addition, since WordNet consists of synonym information on nouns, verbs, adjectives and adverbs, the calculations above are carried out for all possible taggings and the maximum score is used. When one of the words or both do not occur in WordNet, the score is not used. This trades precision for recall in a similar way to Church and Hanks (1990) and Blaheta and Johnson (2001).

5. Results

Each implementation returns a ranked list of word sequences, strongest collocations first (according to the technique). Taking $\mathcal{E}$ to be the set of extracted collocations and $\mathcal{G}$ as the gold standard, recall, r, and precision, p, are defined:

$$r = \frac{|\mathcal{E} \cap \mathcal{G}|}{|\mathcal{G}|} \times 100 \qquad p = \frac{|\mathcal{E} \cap \mathcal{G}|}{|\mathcal{E}|} \times 100$$

As in Evert and Krenn (2001), the results are presented by way of precision and recall graphs with precision and recall calculated for the top N% where $N = 5, 10, 15, \ldots, 100$. For comparison, a baseline was also obtained where scores were allocated at random to each of the bigrams occurring in the corpus. Figures 7 and 8 show recall and precision respectively for each of the techniques. Figure 10 shows the relationships between precision and recall.

Particularly noteworthy is the degree to which church differs from the other curves, especially for precision and recall against precision. This is due to the condition (as described in Section 4.2.) that the score is only calculated if

[6]The mean of the distribution is zero.

For the bigram collocation *pedestrian crossing*, the frequencies and probabilities of alternatives (synonyms) are first calculated:

w	$f(w)$	$p(w_i)$
pedestrian	12	0.60
walker	7	0.35
footer	1	0.05
crossing	8	1.00
crosswalk	0	0.00

Analysing the co-occurrence frequency information yields the following statistics:

P_i	f	p	$\hat{p}$	d_i
pedestrian crossing	18	1	$0.60 \cdot 1 = 0.60$	0.4
pedestrian crosswalk	0	0	$0.60 \cdot 0 = 0$	0
walker crossing	0	0	$0.35 \cdot 1 = 0.35$	−0.35
walker crosswalk	0	0	$0.35 \cdot 0 = 0$	0
footer crossing	0	0	$0.05 \cdot 1 = 0.05$	−0.05
footer crosswalk	0	0	$0.05 \cdot 0 = 0$	0
Total	18	1	1	0

With

$$\sigma(d) = \sqrt{\frac{0.0475}{6}} = 0.22$$

the score for *pedestrian crossing* is then:

$$\frac{0.4}{0.22} = 1.82$$

.

Figure 6: Example for Pearce (2001).

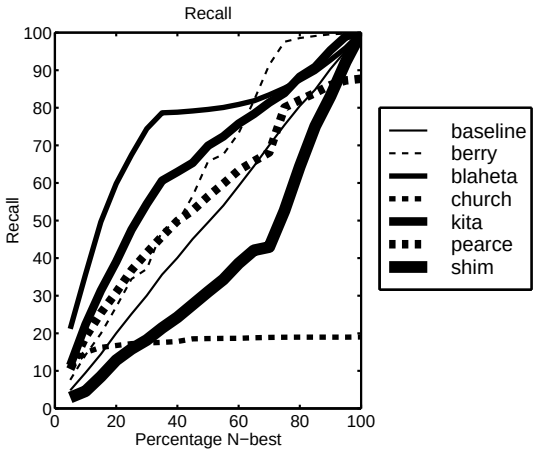

Figure 7: Recall for all techniques.

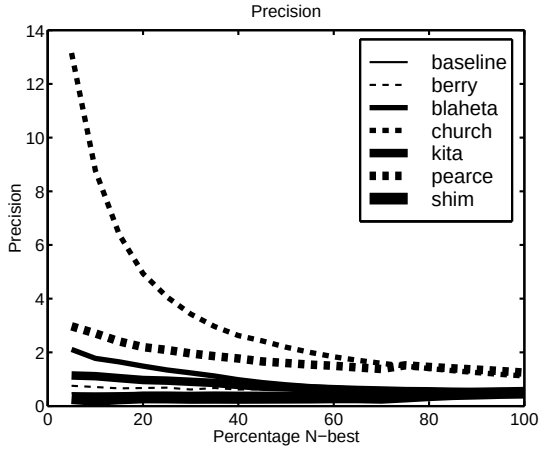

Figure 8: Precision for all techniques.

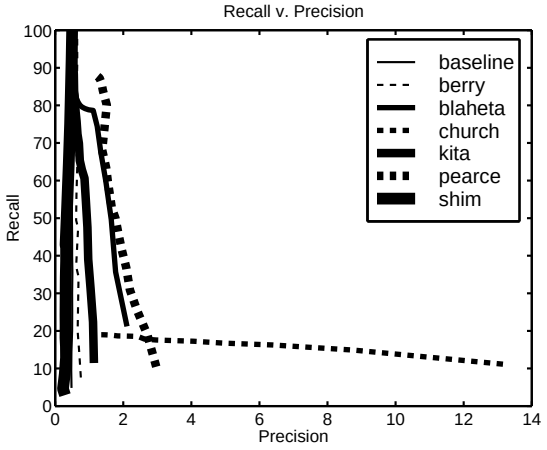

Figure 10: Recall against precision for all techniques.

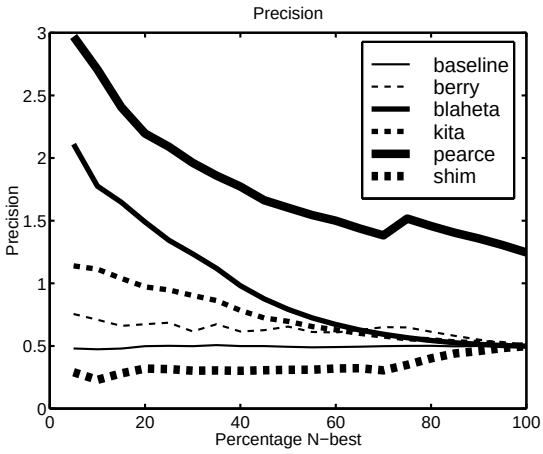

Figure 9: Precision for all techniques except church.

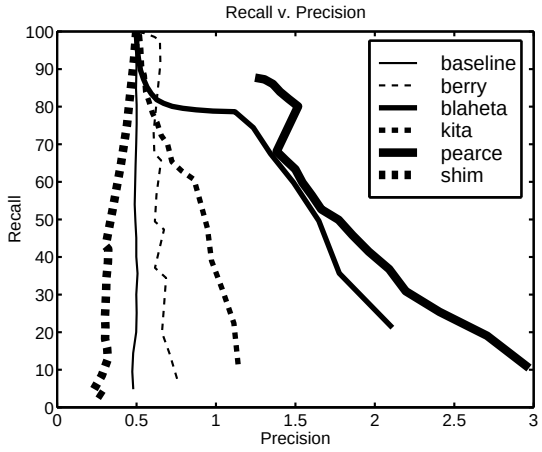

Figure 11: Recall against precision for all techniques except church.

the bigram frequency is greater than 5. Imposing this constraint leads to the loss of over 80% of the collocations in the gold standard since most collocations occur very infrequently. This can be seen in Figure 12 which shows the first few entries in the (proportional) distributions of frequencies in the corpus and in the gold standard. With more data, this problem would be somewhat lessened.

To facilitate discussion of differences between the other techniques, Figures 9 and 11 show the same information as Figures 8 and 10 but without church. The recall curve for blaheta grows particularly smoothly. In contrast, shim performs consistently worse than even the baseline. pearce obtains consistently higher precision than the other techniques (except church) even though it lacks the sense information required. This is possibly due to the fact that by reference to WordNet, bigrams involving closed-class words are not scored. The trade of recall for precision is not nearly so severe as in church since it still extracts nearly 90% of the gold standard.

6. Conclusions

A number of collocation extraction techniques have been evaluated in a bigram collocation extraction task. This task, however, is just one of many that *could* be performed and forms part of an ongoing comparative evaluation of new techniques based on Pearce (2001).

Evaluating against any *one* particular set of criteria is controversial simply because there is no general agreement on the definition of collocation. This is confirmed by the wide range of strategies advanced in the literature as discussed in Section 2. In particular, some of the techniques are specifically designed to extract dependencies with gaps such as those of Berry-Rogghe (1973) and Church and Hanks (1990) although they have been adapted here to process bigram frequency information.

As Krenn and Evert (2001) describe, there is very much a need for the concept of collocation to be precisely defined. However, as has been shown, even in the absence of such conditions, it is still possible for comparative evaluation to be a useful pursuit. Collocation extraction techniques are founded on a variety of assumptions and, in order to fully investigate the utility of a technique, it is necessary to evaluate it against a range of gold standards as well as adopting other strategies such as native speaker judgements and task-based evaluation.

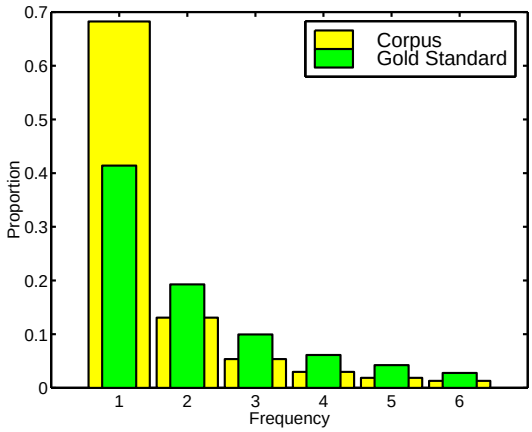

Figure 12: Proportion of bigrams with counts in the range 1 to 6 in the corpus and the gold standard.

Acknowledgements

Much of this work was carried out under a studentship attached to the project 'PSET: Practical Simplification of English Text' funded by the UK EPSRC (ref GR/L53175). Further information about PSET is available at `http://osiris.sunderland.ac.uk/~pset/welcome.html`.

I would also like to thank my supervisor, John Carroll, for his continued support and advice.

7. References

Godelieve L. M. Berry-Rogghe. 1973. The computation of collocations and their relevance to lexical studies. In A. J. Aitken, R. W. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*, pages 103–112. University Press, Edinburgh, New York.

Don Blaheta and Mark Johnson. 2001. Unsupervised learning of multi-word verbs. In *39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39)*, pages 54–60, CNRS - Institut de Recherche en Informatique de Toulouse, and Université des Sciences Sociales, Toulouse, France, July.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, March.

Stefan Evert and Brigitte Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France.

Jean-Philippe Goldman, Luka Nerima, and Eric Wehrli. 2001. Collocation extraction using a syntactic parser. In *39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39)*, pages 61–66, CNRS - Institut de Recherche en Informatique de Toulouse, and Université des Sciences Sociales, Toulouse, France, July.

David C. Howell. 1992. *Statistical Methods For Psychology*. Duxbury Press, Belmont, California, 3rd edition.

Kenji Kita and Hiroaki Ogata. 1997. Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. *Computer Assisted Language Learning: An International Journal*, 10(3):229–238.

Kenji Kita, Yasuhiko Kato, Takashi Omoto, and Yoneo Yano. 1994. A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria. *Journal of Natural Language Processing*, 1(1):21–33.

Brigitte Krenn and Stefan Evert. 2001. Can we do better than frequency? a case study on extracting PP-verb collocations. In *39th Annual Meeting and 10th Conference of the European Chapter of the Association for Computational Linguistics (ACL39)*, pages 39–46, CNRS - Institut de Recherche en Informatique de Toulouse, and Université des Sciences Sociales, Toulouse, France, July.

Dekang Lin. 1998. Extracting collocations from text corpora. In *First Workshop on Computational Terminology*, Montreal, Canada, August.

George A. Miller. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*.

Darren Pearce. 2001. Synonymy in collocation extraction. In *NAACL 2001 Workshop: WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Carnegie Mellon University, Pittsburgh, June.

Judy Pearsall and Patrick Hanks, editors. 1998. *The New Oxford Dictionary of English*. Oxford University Press, August. Machine Readable Version.

Sayori Shimohata, Toshiyuko Sugio, and Junji Nagata. 1997. Retrieving collocations by co-occurrences and word order constraints. In *35th Conference of the Association for Computational Linguistics (ACL'97)*, pages 476–481, Madrid, Spain.

Frank Smadja. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics*, 19(1):143–177, March.