An Extensive Empirical Study of Collocation Extraction Methods

Pavel Pecina
Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic
pecina@ufal.mff.cuni.cz

Abstract

This paper presents a *status quo* of an ongoing research study of collocations – an essential linguistic phenomenon having a wide spectrum of applications in the field of natural language processing. The core of the work is an empirical evaluation of a comprehensive list of automatic collocation extraction methods using precision-recall measures and a proposal of a new approach integrating multiple basic methods and statistical classification. We demonstrate that combining multiple independent techniques leads to a significant performance improvement in comparison with individual basic methods.

1 Introduction and motivation

Natural language cannot be simply reduced to lexicon and syntax. The fact that individual words cannot be combined freely or randomly is common for most natural languages. The ability of a word to combine with other words can be expressed either *intensionally* or *extensionally*. The former case refers to *valency*. Instances of the latter case are called *collocations* (Čermák and Holub, 1982). The term collocation has several other definitions but none of them is widely accepted. Most attempts are based on a characteristic property of collocations: *non-compositionality*. Choueka (1988) defines a collocational expression as "a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components".

The term collocation has both linguistic and lexicographic character. It covers a wide range of lexical phenomena, such as phrasal verbs, light verb compounds, idioms, stock phrases, technological expressions, and proper names. Collocations are of high importance for many applications in the field of NLP. The most desirable ones are machine translation, word sense disambiguation, language generation, and information retrieval. The recent availability of large amounts of textual data has attracted interest in automatic collocation extraction from text. In the last thirty years a number of different methods employing various association measures have been proposed. Overview of the most widely used techniques is given e.g. in (Manning and Schütze, 1999) or (Pearce, 2002). Several researches also attempted to compare existing methods and suggested different evaluation schemes, e.g Kita (1994) or Evert (2001). A comprehensive study of statistical aspects of word cooccurrences can be found in (Evert, 2004).

In this paper we present a compendium of 84 methods for automatic collocation extraction. They came from different research areas and some of them have not been used for this purpose yet. A brief overview of these methods is followed by their comparative evaluation against manually annotated data by the means of precision and recall measures. In the end we propose a statistical classification method for combining multiple methods and demonstrate a substantial performance improvement.

In our research we focus on two-word (*bigram*) collocations, mainly for the reason that experiments with longer expressions would require processing of much larger amounts of data and limited scalability of some methods to high order n-grams. The experiments are performed on Czech data.

Proceedings of the ACL Student Research Workshop, pages 13–18,
Ann Arbor, Michigan, June 2005.

2 Collocation extraction

Most methods for collocation extraction are based on verification of typical collocation properties. These properties are formally described by mathematical formulas that determine the degree of association between components of collocation. Such formulas are called *association measures* and compute an *association score* for each collocation candidate extracted from a corpus. The scores indicate a chance of a candidate to be a collocation. They can be used for ranking or for classification – by setting a threshold. Finding such a threshold depends on the intended application.

The most widely tested property of collocations is *non-compositionality*: If words occur together more often than by a chance, then this is the evidence that they have a special function that is not simply explained as a result of their combination (Manning and Schütze, 1999). We think of a corpus as a randomly generated sequence of words that is viewed as a sequence of word pairs. Occurrence frequencies of these bigrams are extracted and kept in contingency tables (Table 1a). Values from these tables are used in several association measures that reflect how much the word coocurrence is accidental. A list of such measures is given in Table 2 and includes: estimation of bigram and unigram probabilities (rows 3–5), mutual information and derived measures (6–11), statistical tests of independence (12–16), likelihood measures (17–18), and various other heuristic association measures and coefficients (19–57).

Another frequently tested property is taken directly from the definition that a collocation is a *syntactic and semantic unit*. For each bigram occurring in the corpus, information of its *empirical context* (frequencies of open-class words occurring within a specified context window) and left and right *immediate contexts* (frequencies of words immediately preceding or following the bigram) is extracted (Table 1b). By determining the entropy of the immediate contexts of a word sequence, the association measures rank collocations according to the assumption that they occur as units in a (information-theoretically) noisy environment (Shimohata et al., 1997) (58–62). By comparing empirical contexts of a word sequence and its components, the association measures rank collocations according to the assumption that semantically non-compositional expressions typically occur in different contexts than their components (Zhai, 1997). Measures (63–76) have information theory background and measures (77–84) are adopted from the field of information retrieval. Context association measures are mainly used for extracting idioms.

a)

$a = f(xy)$	$b = f(x\bar{y})$	$f(x*)$
$c = f(\bar{x}y)$	$d = f(\bar{x}\bar{y})$	$f(\bar{x}*)$
$f(*y)$	$f(*\bar{y})$	N

b)

C_w	empirical context of w
C_{xy}	empirical context of xy
C^l_{xy}	left immediate context of xy
C^r_{xy}	right immediate context of xy

Table 1: a) A contingency table with observed frequencies and marginal frequencies for a bigram xy; $\bar{w}$ stands for any word except w; $*$ stands for any word; N is a total number of bigrams. The table cells are sometimes referred as f_{ij}. Statistical tests of independence work with contingency tables of expected frequencies $\hat{f}(xy) = f(x*)f(*y)/N$. b) Different notions of empirical contexts.

Besides all the association measures described above, we also take into account other recommended measures (1–2) (Manning and Schütze, 1999) and some basic linguistic characteristics used for filtering non-collocations (85–87). This information can be obtained automatically from morphological taggers and syntactic parsers available with reasonably high accuracy for many languages.

3 Empirical evaluation

Evaluation of collocation extraction methods is a complicated task. On one hand, different applications require different setting of association score thresholds. On the other hand, methods give different results within different ranges of their association scores. We need a complex evaluation scheme covering all demands. In such a case, Evert (2001) and other authors suggest using *precision* and *recall* measures on a full reference data or on *n-best* lists.

Data. All the presented experiments were performed on morphologically and syntactically annotated Czech text from the *Prague Dependency Treebank* (PDT) (Hajič et al., 2001). Dependency trees were broken down into *dependency bigrams* consisting of: *lemmas* and *part-of-speech* of the components, and *type of dependence* between the components.

For each bigram type we counted frequencies in its contingency table, extracted empirical and immediate contexts, and computed all the 84 association measures from Table 2. We processed 81 614 sen-

#	Name	Formula
	1. **Mean component offset**	$\frac{1}{n}\sum_{i=1}^{n} d_i$
	2. **Variance component offset**	$\frac{1}{n-1}\sum_{i=1}^{n}(d_i-\bar{d})^2$
	3. **Joint probability**	$P(xy)$
	4. **Conditional probability**	$P(y\|x)$
	5. **Reverse conditional prob.**	$P(x\|y)$
⋆	6. **Pointwise mutual inform.**	$\log\frac{P(xy)}{P(x*)P(*y)}$
	7. **Mutual dependency (*MD*)**	$\log\frac{P(xy)^2}{P(x*)P(*y)}$
	8. **Log frequency biased *MD***	$\log\frac{P(xy)^2}{P(x*)P(*y)}+\log P(xy)$
	9. **Normalized expectation**	$\frac{2f(xy)}{f(x*)+f(*y)}$
⋆	10. **Mutual expectation**	$\frac{2f(xy)}{f(x*)+f(*y)}\cdot P(xy)$
	11. **Salience**	$\log\frac{P(xy)^2}{P(x*)P(*y)}\cdot\log f(xy)$
	12. **Pearson's χ^2 test**	$\sum_{ij}\frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
	13. **Fisher's exact test**	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
	14. **t test**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
	15. **z score**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
	16. **Poison significance measure**	$\frac{\hat{f}(xy)-f(xy)\log\hat{f}(xy)+\log f(xy)!}{\log N}$
	17. **Log likelihood ratio**	$-2\sum_{ij} f_{ij}\log\frac{f_{ij}}{\hat{f}_{ij}}$
	18. **Squared log likelihood ratio**	$-2\sum_{ij}\frac{\log f_{ij}^2}{\hat{f}_{ij}}$
	Association coefficients:	
	19. **Russel-Rao**	$\frac{a}{a+b+c+d}$
	20. **Sokal-Michiner**	$\frac{a+d}{a+b+c+d}$
⋆	21. **Rogers-Tanimoto**	$\frac{a+d}{a+2b+2c+d}$
	22. **Hamann**	$\frac{(a+d)-(b+c)}{a+b+c+d}$
	23. **Third Sokal-Sneath**	$\frac{b+c}{a+d}$
	24. **Jaccard**	$\frac{a}{a+b+c}$
⋆	25. **First Kulczynsky**	$\frac{a}{b+c}$
	26. **Second Sokal-Sneath**	$\frac{a}{a+2(b+c)}$
	27. **Second Kulczynski**	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
	28. **Fourth Sokal-Sneath**	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$
	29. **Odds ratio**	$\frac{ad}{bc}$
	30. **Yulle's ω**	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
⋆	31. **Yulle's Q**	$\frac{ad-bc}{ad+bc}$
	32. **Driver-Kroeber**	$\frac{a}{\sqrt{(a+b)(a+c)}}$
	33. **Fifth Sokal-Sneath**	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
	34. **Pearson**	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
	35. **Baroni-Urbani**	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
	36. **Braun-Blanquet**	$\frac{a}{\max(a+b,a+c)}$
	37. **Simpson**	$\frac{a}{\min(a+b,a+c)}$
	38. **Michael**	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
	39. **Mountford**	$\frac{2a}{2bc+ab+ac}$
	40. **Fager**	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2}\max(b,c)$
	41. **Unigram subtuples**	$\log\frac{ad}{bc}-3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
	42. ***U* cost**	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$
	43. ***S* cost**	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$
	44. ***R* cost**	$\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$
	45. ***T* combined cost**	$\sqrt{U\times S\times R}$
	46. **Phi**	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
	47. **Kappa**	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
	48. ***J* measure**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$

#	Name	Formula
	49. **Gini index**	$\max[P(x*)(P(y\|x)^2+P(\bar{y}\|x)^2)-P(*y)^2$ $+P(\bar{x*})(P(y\|\bar{x})^2+P(\bar{y}\|\bar{x})^2)-P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2$ $+P(*\bar{y})(P(x\|\bar{y})^2+P(\bar{x}\|\bar{y})^2)-P(\bar{x}*)^2]$
	50. **Confidence**	$\max[P(y\|x),P(x\|y)]$
	51. **Laplace**	$\max[\frac{NP(xy)+1}{NP(x*)+2},\frac{NP(xy)+1}{NP(*y)+2}]$
	52. **Conviction**	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})},\frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
	53. **Piatersky-Shapiro**	$P(xy)-P(x*)P(*y)$
	54. **Certainity factor**	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)},\frac{P(x\|y)-P(x*)}{1-P(x*)}]$
	55. **Added value (*AV*)**	$\max[P(y\|x)-P(*y),P(x\|y)-P(x*)]$
⋆	56. **Collective strength**	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)}\cdot$ $\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
	57. **Klosgen**	$\sqrt{P(xy)}\cdot AV$
	Context measures:	
⋆	58. **Context entropy**	$-\sum_w P(w\|C_{xy})\log P(w\|C_{xy})$
	59. **Left context entropy**	$-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
	60. **Right context entropy**	$-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
⋆	61. **Left context divergence**	$P(x*)\log P(x*)$ $-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
	62. **Right context divergence**	$P(*y)\log P(*y)$ $-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
	63. **Cross entropy**	$-\sum_w P(w\|C_x)\log P(w\|C_y)$
	64. **Reverse cross entropy**	$-\sum_w P(w\|C_y)\log P(w\|C_x)$
	65. **Intersection measure**	$\frac{2\|C_x\cap C_y\|}{\|C_x\|+\|C_y\|}$
	66. **Euclidean norm**	$\sqrt{\sum_w(P(w\|C_x)-P(w\|C_y))^2}$
	67. **Cosine norm**	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2\cdot\sum_w P(w\|C_y)^2}$
	68. ***L1* norm**	$\sum_w\|P(w\|C_x)-P(w\|C_y)\|$
	69. **Confusion probability**	$\sum_w\frac{P(x\|C_w)P(y\|C_w)P(w)}{P(x*)}$
	70. **Reverse confusion prob.**	$\sum_w\frac{P(y\|C_w)P(x\|C_w)P(w)}{P(*y)}$
⋆	71. **Jensen-Shannon diverg.**	$\frac{1}{2}[D(p(w\|C_x)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))$ $+D(p(w\|C_y)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))]$
	72. **Cosine of pointwise *MI***	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2}\cdot\sqrt{\sum_w MI(w,y)^2}}$
⋆	73. **KL divergence**	$\sum_w P(w\|C_x)\log\frac{P(w\|C_x)}{P(w\|C_y)}$
⋆	74. **Reverse KL divergence**	$\sum_w P(w\|C_y)\log\frac{P(w\|C_y)}{P(w\|C_x)}$
	75. **Skew divergence**	$D(p(w\|C_x)\|\|\alpha(w\|C_y)+(1-\alpha)p(w\|C_x))$
	76. **Reverse skew divergence**	$D(p(w\|C_y)\|\|\alpha p(w\|C_x)+(1-\alpha)p(w\|C_y))$
	77. **Phrase word coocurrence**	$\frac{1}{2}(\frac{f(x\|C_{xy})}{f(xy)}+\frac{f(y\|C_{xy})}{f(xy)})$
	78. **Word association**	$\frac{1}{2}(\frac{f(x\|C_y)-f(xy)}{f(xy)}+\frac{f(y\|C_x)-f(xy)}{f(xy)})$
	Cosine context similarity:	$\frac{1}{2}(\cos(\mathbf{c}_x,\mathbf{c}_{xy})+\cos(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \cos(\mathbf{c}_x,\mathbf{c}_y)=\frac{\sum x_i y_i}{\sqrt{\sum {x_i}^2}\cdot\sqrt{\sum {y_i}^2}}$
⋆	79. **in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
	80. **in *tf* vector space**	$z_i=f(w_i\|C_z)$
	81. **in *tf*$\cdot$*idf* vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)}$; $df(w_i)=\|\{x:w_i\epsilon C_x\}\|$
	Dice context similarity:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_i y_i}{\sum {x_i}^2+\sum {y_i}^2}$
⋆	82. **in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
⋆	83. **in *tf* vector space**	$z_i=f(w_i\|C_z)$
⋆	84. **in *tf*$\cdot$*idf* vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)}$; $df(w_i)=\|\{x:w_i\epsilon C_x\}\|$
	Linguistic features:	
⋆	85. **Part of speech**	{*Adjective:Noun, Noun:Noun, Noun:Verb, …*}
⋆	86. **Dependency type**	{*Attribute, Object, Subject, …*}
	87. **Dependency structure**	{↗, ↖}

Table 2: Association measures and linguistic features used in bigram collocation extraction methods. ⋆ denotes those selected by the attribute selection method discussed in Section 4. References can be found at the end of the paper.

tences with 1 255 590 words and obtained a total of 202 171 different dependency bigrams.

Krenn (2000) argues that collocation extraction methods should be evaluated against a reference set of collocations manually extracted from the full candidate data from a corpus. However, we reduced the full candidate data from PDT to 21 597 bigram by filtering out any bigrams which occurred 5 or less times in the data and thus we obtained a *reference data set* which fulfills requirements of a sufficient size and a minimal frequency of observations which is needed for the assumption of normal distribution required by some methods.

We manually processed the entire reference data set and extracted bigrams that were considered to be collocations. At this point we applied *part-of-speech* filtering: First, we identified *POS patterns* that never form a collocation. Second, all dependency bigrams having such a *POS* pattern were removed from the reference data and a final reference set of 8 904 bigrams was created. We no longer consider bigrams with such patterns to be collocation candidates.

This data set contained 2 649 items considered to be collocations. The a priori probability of a bigram to be a collocation was 29.75 %. A stratified one-third subsample of this data was selected as *test data* and used for evaluation and testing purposes in this work. The rest was taken apart and used as *training data* in later experiments.

Evaluation metrics. Since we manually annotated the entire reference data set we could use the suggested *precision* and *recall* measures (and their harmonic mean *F-measure*). A collocation extraction method using any association measure with a given threshold can be considered a classifier and the measures can be computed in the following way:

$$Precision = \frac{\#\ correctly\ classified\ collocations}{\#\ total\ predicted\ as\ collocations}$$

$$Recall = \frac{\#\ correctly\ classified\ collocations}{\#\ total\ collocations}$$

The higher these scores, the better the classifier is. By changing the threshold we can tune the classifier performance and "trade" recall for precision. Therefore, collocation extraction methods can be thoroughly compared by comparing their *precision--recall curves*: The closer the curve to the top right corner, the better the method is.

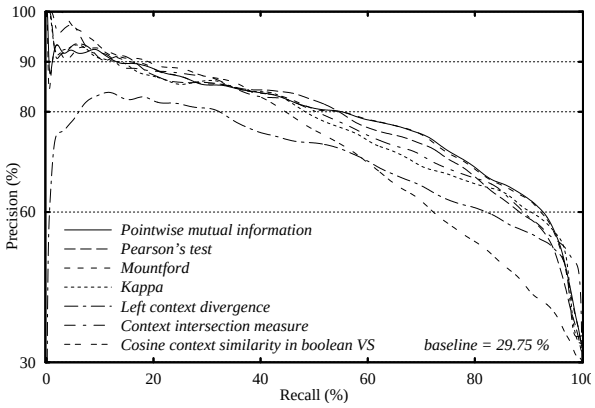

Figure 1: Precision-recall curves for selected assoc. measures.

Results. Presenting individual results for all of the 84 association measures is not possible in a paper of this length. Therefore, we present precision-recall graphs only for the best methods from each group mentioned in Section 2; see Figure 1. The baseline system that classifies bigrams randomly, operates with a precision of 29.75 %. The overall best result was achieved by *Pointwise mutual information*: 30 % recall with 85.5 % precision (F-measure 44.4), 60 % recall with 78.4 % precision (F-measure 68.0), and 90 % recall with 62.5 % precision (F-measure 73.8).

4 Statistical classification

In the previous section we mentioned that collocation extraction is a classification problem. Each method classifies instances of the candidate data set according to the values of an association score. Now we have several association scores for each candidate bigram and want to combine them together to achieve better performance. A motivating example is depicted in Figure 3: Association scores of *Pointwise mutual information* and *Cosine context similarity* are independent enough to be linearly combined to provide better results. Considering all association measures, we deal with a problem of high-dimensional classification into two classes.

In our case, each bigram x is described by the *attribute vector* $\mathbf{x}=(x_1,\ldots,x_{87})$ consisting of linguistic features and association scores from Table 2. Now we look for a function assigning each bigram one class : $f(\mathbf{x}) \rightarrow$*{collocation, non-collocation}*. The result of this approach is similar to setting a threshold of the association score in methods us-

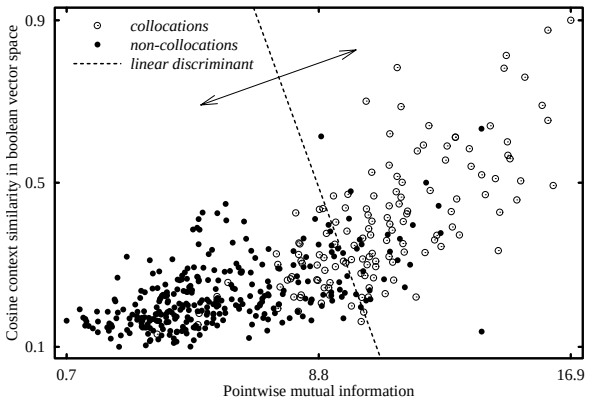

Figure 2: Data visualization in two dimensions. The dashed line denotes a linear discriminant obtained by logistic linear regression. By moving this boundary we can tune the classifier output (a 5 % stratified sample of the test data is displayed).

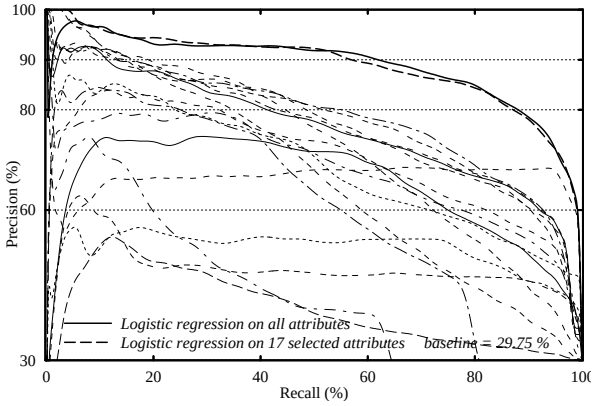

Figure 3: Precision-recall curves of two classifiers based on i) logistic linear regression on the full set of 87 attributes and ii) on the selected subset with 17 attributes. The thin unlabeled curves refer to the methods from the 17 selected attributes

ing one association measure, which is not very usefull for our purpose. Some classification methods, however, output also the predicted probability $P(x\,is\,collocation)$ that can be considered a regular association measure as described above. Thus, the classification method can be also tuned by changing a threshold of this probability and can be compared with other methods by the same means of precision and recall.

One of the basic classification methods that gives a predicted probability is *Logistic linear regression*. The model defines the predicted probability as:

$$P(x\,is\,collocation) = \frac{\exp^{\beta_0+\beta_1 x_1 ... +\beta_n x_n}}{1+\exp^{\beta_0+\beta_1 x_1 ... +\beta_n x_n}}$$

where the coefficients β_i are obtained by the *iteratively reweighted least squares* (IRLS) algorithm which solves the weighted least squares problem at each iteration. Categorial attributes need to be transformed to numeric *dummy variables*. It is also recommended to *normalize* all numeric attributes to have zero mean and unit variance.

We employed the datamining software *Weka* by Witten and Frank (2000) in our experiments. As *training data* we used a two-third subsample of the reference data described above. The *test data* was the same as in the evaluation of the basic methods.

By combining all the 87 attributes, we achieved the results displayed in Table 3 and illustrated in Figure 3. At a recall level of 90 % the relative increase in precision was 35.2 % and at a precision level of 90 % the relative increase in recall was impressive 242.3 %.

Attribute selection. In the final step of our experiments, we attempted to reduce the attribute space of our data and thus obtain an attribute subset with the same prediction ability. We employed a *greedy stepwise* search method with attribute subset evaluation via logistic regression implemented in Weka. It performs a greedy search through the space of attribute subsets and iteratively merges subsets that give the best results until the performance is no longer improved.

We ended up with a subset consisting of the following 17 attributes: (6, 10, 21, 25, 31, 56, 58, 61, 71, 73, 74, 79, 82, 83, 84, 85, 86) which are also marked in Table 2. The overview of achieved results is shown in Table 3 and precision-recall graphs of the selected attributes and their combinations are in Figure 3.

5 Conclusions and future work

We implemented 84 automatic collocation extraction methods and performed series of experiments on morphologically and syntactically annotated data. The methods were evaluated against a reference set of collocations manually extracted from the

	Recall			*Precision*		
	30	60	90	70	80	90
P. mutual information	85.5	78.4	62.5	78.0	56.0	16.3
Logistic regression-17	92.6	89.5	84.5	96.7	86.7	55.8
Absolute improvement	7.1	11.1	**22.0**	17.7	30.7	**39.2**
Relative improvement	8.3	14.2	**35.2**	23.9	54.8	**242.3**

Table 3: Precision (the 3 left columns) and recall (the 3 right columns) scores (in %) for the best individual method and linear combination of the 17 selected ones.

same source. The best method (*Pointwise mutual information*) achieved 68.3 % recall with 73.0 % precision (F-measure 70.6) on this data. We proposed to combine the association scores of each candidate bigram and employed *Logistic linear regression* to find a linear combination of the association scores of all the basic methods. Thus we constructed a collocation extraction method which achieved 80.8 % recall with 84.8 % precision (F-measure 82.8). Furthermore, we applied an attribute selection technique in order to lower the high dimensionality of the classification problem and reduced the number of regressors from 87 to 17 with comparable performance. This result can be viewed as a kind of evaluation of basic collocation extraction techniques. We can obtain the smallest subset that still gives the best result. The other measures therefore become uninteresting and need not be further processed and evaluated.

The reseach presented in this paper is in progress. The list of collocation extraction methods and association measures is far from complete. Our long term goal is to collect, implement, and evaluate all available methods suitable for this task, and release the toolkit for public use.

In the future, we will focus especially on improving quality of the training and testing data, employing other classification and attribute-selection techniques, and performing experiments on English data. A necessary part of the work will be a rigorous theoretical study of all applied methods and appropriateness of their usage. Finally, we will attempt to demonstrate contribution of collocations in selected application areas, such as machine translation or information retrieval.

Acknowledgments

This research has been supported by the Ministry of Education of the Czech Republic, project MSM 0021620838. I would also like to thank my advisor, Dr. Jan Hajič, for his continued support.

References

Y. Choueka. 1988. Looking for needles in a haystack or locating interesting collocational expressions in large textual databases. In *Proceedings of the RIAO*, pages 43–38.

I. Dagan, L. Lee, and F. Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning*, 34.

T. E. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

S. Evert and B. Krenn. 2001. Methods for the qualitative evaluation of lexical association measures. In *Proceedings 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195.

S. Evert. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.

J. Hajič, E. Hajičová, P. Pajas, J. Panevová, P. Sgall, and B. Vidová-Hladká. 2001. Prague dependency treebank 1.0. Published by LDC, University of Pennsylvania.

K. Kita, Y. Kato, T. Omoto, and Y. 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.

B. Krenn. 2000. Collocation Mining: Exploiting Corpora for Collocation Idenfication and Representation. In *Proceedings of KONVENS 2000*.

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

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

D. Pearce. 2002. A comparative evaluation of collocation extraction techniques. In *Third International Conference on language Resources and Evaluation*, Las Palmas, Spain.

T. Pedersen. 1996. Fishing for exactness. In *Proceedings of the South Central SAS User's Group Conference*, pages 188–200, Austin, TX.

S. Shimohata, T. Sugio, and J. Nagata. 1997. Retrieving collocations by co-occurrences and word order constraints. In *Proc. of the 35th Annual Meeting of the ACL and 8th Conference of the EACL*, pages 476–81, Madrid. Spain.

P. Tan, V. Kumar, and J. Srivastava. 2002. Selecting the right interestingness measure for association patterns. In *Proceedings of the Eight A CM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

A. Thanopoulos, N. Fakotakis, and G. Kokkinakis. 2002. Comparative evaluation of collocation extraction metrics. In *3rd International Conference on Language Resources and Evaluation*, volume 2, pages 620–625, Las Palmas, Spain.

F. Čermák and J. Holub. 1982. *Syntagmatika a paradigmatika česk eho slova: Valence a kolokabilita*. Státní pedagogické nakladatelství, Praha.

I. H. Witten and E. Frank. 2000. *Data Mining: Practical machine learning tools with Java implementations*. Morgan Kaufmann, San Francisco.

C. Zhai. 1997. Exploiting context to identify lexical atoms – A statistical view of linguistic context. In *International and Interdisciplinary Conference on Modelling and Using Context (CONTEXT-97)*.