Appendix

Table 3 The inventory of lexical association measures used for collocation extraction used in our experiments

#	*Name*	*Formula*
1.	**Joint probability**	$P(xy)$
2.	**Conditional probability**	$P(y\|x)$
3.	**Reverse conditional probability**	$P(x\|y)$
4.	**Pointwise mutual information**	$\log \frac{P(xy)}{P(x*)P(*y)}$
5.	**Mutual dependency (*MD*)**	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
*6.	**Log frequency biased *MD***	$\log \frac{P(xy)^2}{P(x*)P(*y)} + \log P(xy)$
7.	**Normalized expectation**	$\frac{2f(xy)}{f(x*)+f(*y)}$
8.	**Mutual expectation**	$\frac{2f(xy)}{f(x*)+f(*y)} \cdot P(xy)$
9.	**Salience**	$\log \frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$
10.	**Pearson's χ^2 test**	$\sum_{i,j} \frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
11.	**Fisher's exact test**	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$
12.	**t test**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
13.	**z score**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
14.	**Poison significance measure**	$\frac{\hat{f}(xy)-f(xy)\log\hat{f}(xy)+\log f(xy)!}{\log N}$
15.	**Log likelihood ratio**	$-2\sum_{i,j} f_{ij} \log \frac{f_{ij}}{\hat{f}_{ij}}$
16.	**Squared log likelihood ratio**	$-2\sum_{i,j} \frac{\log f_{ij}^2}{\hat{f}_{ij}}$
Association coefficients:		
17.	**Russel-Rao**	$\frac{a}{a+b+c+d}$
18.	**Sokal-Michiner**	$\frac{a+d}{a+b+c+d}$
19.	**Rogers-Tanimoto**	$\frac{a+d}{a+2b+2c+d}$
20.	**Hamann**	$\frac{(a+d)-(b+c)}{a+b+c+d}$
21.	**Third Sokal-Sneath**	$\frac{b+c}{a+d}$
22.	**Jaccard**	$\frac{a}{a+b+c}$
*23.	**First Kulczynsky**	$\frac{a}{b+c}$
24.	**Second Sokal-Sneath**	$\frac{a}{a+2(b+c)}$
25.	**Second Kulczynski**	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
26.	**Fourth Sokal-Sneath**	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$
27.	**Odds ratio**	$\frac{ad}{bc}$
28.	**Yulle's ω**	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
29.	**Yulle's Q**	$\frac{ad-bc}{ad+bc}$
30.	**Driver-Kroeber**	$\frac{a}{\sqrt{(a+b)(a+c)}}$
31.	**Fifth Sokal-Sneath**	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$

Table 3 continued

#	*Name*	*Formula*
32.	**Pearson**	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
33.	**Baroni-Urbani**	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
34.	**Braun-Blanquet**	$\frac{a}{\max(a+b,a+c)}$
35.	**Simpson**	$\frac{a}{\min(a+b,a+c)}$
36.	**Michael**	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
37.	**Mountford**	$\frac{2a}{2bc+ab+ac}$
38.	**Fager**	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2}\max(b,c)$
⋆39.	**Unigram subtuples**	$\log\frac{ad}{bc}-3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
40.	***U* cost**	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$
⋆41.	***S* cost**	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$
42.	***R* cost**	$\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$
43.	***T* combined cost**	$\sqrt{U\times S\times R}$
44.	**Phi**	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
45.	**Kappa**	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
46.	***J* measure**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
47.	**Gini index**	$\max[P(x*)(P(y\|x)^2+P(\bar{y}\|x)^2)-P(*y)^2$ $+P(\bar{x}*)(P(y\|\bar{x})^2+P(\bar{y}\|\bar{x})^2)-P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2$ $+P(*\bar{y})(P(x\|\bar{y})^2+P(\bar{x}\|\bar{y})^2)-P(\bar{x}*)^2]$
48.	**Confidence**	$\max[P(y\|x),P(x\|y)]$
49.	**Laplace**	$\max[\frac{NP(xy)+1}{NP(x*)+2},\frac{NP(xy)+1}{NP(*y)+2}]$
50.	**Conviction**	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})},\frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
51.	**Piatersky-Shapiro**	$P(xy)-P(x*)P(*y)$
52.	**Certainity factor**	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)},\frac{P(x\|y)-P(x*)}{1-P(x*)}]$
53.	**Added value (*AV*)**	$\max[P(y\|x)-P(*y),P(x\|y)-P(x*)]$
54.	**Collective strength**	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)}\cdot$ $\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
55.	**Klosgen**	$\sqrt{P(xy)}\cdot AV$
Context measures:		
56.	**Context entropy**	$-\sum_w P(w\|C_{xy})\log P(w\|C_{xy})$
⋆57.	**Left context entropy**	$-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
⋆58.	**Right context entropy**	$-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
⋆59.	**Left context divergence**	$P(x*)\log P(x*)-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
60.	**Right context divergence**	$P(*y)\log P(*y)-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$

Table 3 continued

#	*Name*	*Formula*
61.	**Cross entropy**	$-\sum_w P(w\|C_x)\log P(w\|C_y)$
⋆62.	**Reverse cross entropy**	$-\sum_w P(w\|C_y)\log P(w\|C_x)$
63.	**Intersection measure**	$\frac{2\|C_x \cap C_y\|}{\|C_x\|+\|C_y\|}$
64.	**Euclidean norm**	$\sqrt{\sum_w (P(w\|C_x) - P(w\|C_y))^2}$
65.	**Cosine norm**	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2 \cdot \sum_w P(w\|C_y)^2}$
66.	***L1* norm**	$\sum_w \|P(w\|C_x) - P(w\|C_y)\|$
67.	**Confusion probability**	$\sum_w \frac{P(x\|C_w)P(y\|C_w)P(w)}{P(x*)}$
⋆68.	**Reverse confusion probability**	$\sum_w \frac{P(y\|C_w)P(x\|C_w)P(w)}{P(*y)}$
69.	**Jensen-Shannon divergence**	$\frac{1}{2}[D(p(w\|C_x)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))$
		$+D(p(w\|C_y)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))]$
70.	**Cosine of pointfwise *MI***	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2} \cdot \sqrt{\sum_w MI(w,y)^2}}$
71.	**KL divergence**	$\sum_w P(w\|C_x)\log\frac{P(w\|C_x)}{P(w\|C_y)}$
72.	**Reverse KL divergence**	$\sum_w P(w\|C_y)\log\frac{P(w\|C_y)}{P(w\|C_x)}$
73.	**Skew divergence**	$D(p(w\|C_x)\|\|\alpha p(w\|C_y)+(1-\alpha)p(w\|C_x))$
74.	**Reverse skew divergence**	$D(p(w\|C_y)\|\|\alpha p(w\|C_x)+(1-\alpha)p(w\|C_y))$
⋆75.	**Phrase word coocurrence**	$\frac{1}{2}(\frac{f(x\|C_{xy})}{f(xy)}+\frac{f(y\|C_{xy})}{f(xy)})$
76.	**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:	$\frac{1}{2}(\cos(\mathbf{c}_x,\mathbf{c}_{xy})+\cos(\mathbf{c}_y,\mathbf{c}_{xy}))$
		$\mathbf{c}_z=(z_i);\ \cos(\mathbf{c}_x,\mathbf{c}_y)=\frac{\sum x_i y_i}{\sqrt{\sum x_i{}^2}\cdot\sqrt{\sum y_i{}^2}}$
⋆77.	**in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
78.	**in tf vector space**	$z_i=f(w_i\|C_z)$
79.	**in $tf \cdot idf$ vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x: w_i \varepsilon C_x\}\|$
	Dice context similarity:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$
		$\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_i y_i}{\sum x_i{}^2+\sum y_i{}^2}$
80.	**in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
⋆81.	**in tf vector space**	$z_i=f(w_i\|C_z)$
⋆82.	**in $tf \cdot idf$ vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x: w_i \varepsilon C_x\}\|$

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

C_w	empirical context of w
C_{xy}	empirical context of xy
C_{xy}^l	left immediate context of xy
C_{xy}^r	right immediate context of xy

A contingency table contains observed joint 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 to as f_{ij}. Statistical tests of independence work with contingency tables of expected frequencies $\hat{f}(xy)=f(x*)f(*y)/N$