An Extensive Empirical Study of Collocation Extraction Methods

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Charles University, Prague

June 27, 2005

Outline

Outline

Definitions I

Firth (1951):

> *"Collocations of a given word are statements of the habitual or customary places of that word."*

Choueka (1988):

> *"A collocation is a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components."*

Čermák (1982):

> *"Individual words cannot be combined freely or randomly only by syntactic rules. The ability of a word to combine with other words (collocability) can be expressed:*
>
> *a) intensionally* $\rightarrow$ *valency*
> *b) extensionally"* $\rightarrow$ *collocations*

Characteristic Properties

Non-compositionality *(kick the bucket, carriage return, white man)*

- *The meaning of a collocation is not a straightforward composition of the meaning of its parts.*

Non-substitutability *(yellow wine, hit the bucket, make homework)*

- *Components of collocation cannot be substituted with a related word or a synonym.*

Non-modifiability *(give a big hand, poor as church mice)*

- *Collocations cannot be modified or syntactically transformed.*

Other properties

- *Collocations are not necessarily adjacent.* *(knock the door)*
- *Collocations cannot be directly translated.* *(ice cream)*
- *Collocations are domain-specific.* *(carriage return)*
- *Judging collocations is subjective.* *(new company)*

Types of Collocations

Collocations have both linguistic and lexicographic character and covers a wide range of lexical phenomena:

- light verb compounds – *verbs with little semantic content* *(take, make,do)*
- verb particle constructions, phrasal verbs *(look up, take off, tell off)*
- idioms – *fixed phrases* *(kick the bucket)*
- stock phrases *(good morning)*
- technological expresions – *concepts or objects in tech. dom.* *(hard disk)*
- proper names *(Ann Arbor)*

Motivation

Collocations can be used in a wide range of fields:

- Lexicography
- Machine translation
- Information retrieval, information extraction
- Word sense disambiguation
- Spell/grammar/style-checking
- Text classification and summarization
- Keyword extraction
- Language modeling
- Language generation

The Tasks

To build a collocation lexicon.

1. Creating manually annotated reference data
   - *of reasonable size.*
2. Evaluation of collocation extraction methods
   - *interval-wise by the means of precision-recall.*
3. Combining association measures for collocation extraction
   - *and achieve "better" results.*
4. Reduce number of combined measures
   - *and select the "best subset" of available association measures.*

Focus on bigram collocations

1. *Processing of longer expressions requires larger amounts of data.*
2. *Scalability of some methods to high order n-grams is limited.*

Outline

Collocation Extraction

- Most methods are based on verification of typical collocation properties.
- These properties are formally described by mathematical formulas that determine degree of association between words.
- Such formulas are called association measures and compute association score for each collocation candidate from a corpus.
- The scores indicate a chance of a candidate to be a collocation.
- The scores can be used for ranking or for classification:

Ranking

red cross	*15.66*
decimal point	*14.01*
arithmetic operation	*10.52*
paper feeder	*10.17*
system type	*3.54*
and others	*0.54*
program in	*0.35*
level is	*0.25*

Classification

red cross	*1*
decimal point	*1*
arithmetic operation	*1*
paper feeder	*1*
system type	*0*
and others	*0*
program in	*0*
level is	*0*

The Methodology

1. Identifying Word Base Forms:
   - Surface forms
   - Stems or lemmas
   - Lemmas with additional morphosyntactic features
2. Extracting all possible collocation candidates:
   - Consequent word n-grams (*multi-word expressions*)
   - Sliding window
   - Syntactic structures (*dependency n-grams*)
3. Collecting coocurrence statistics:
   - Frequency of word and n-gram occurrences
   - Immediate contexts
   - Global contexts
4. Computing association measures
5. Ranking or classification

Word Base Forms

Problem:

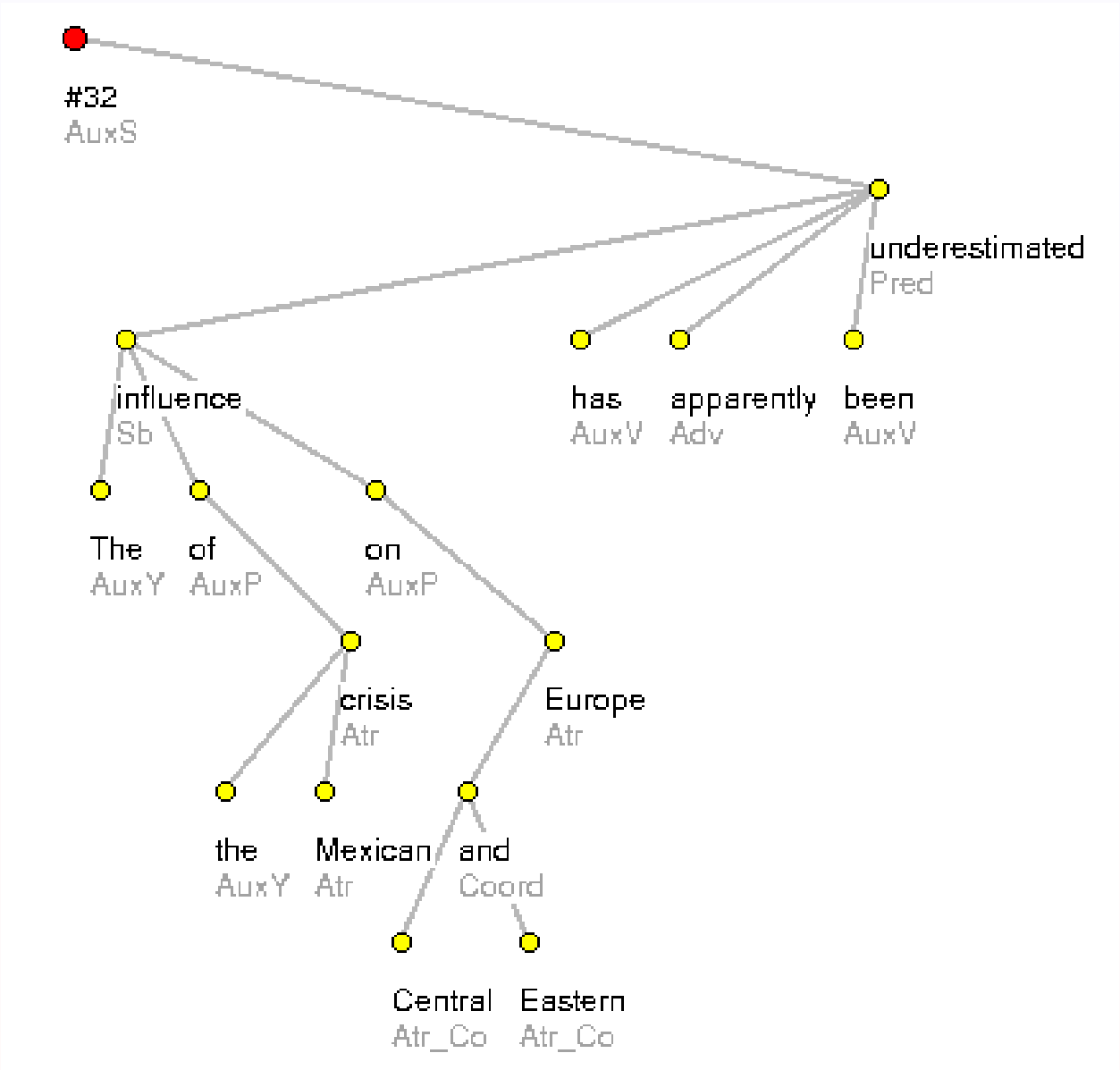
- Surface word forms too specific (*rich morphology, we work with Czech*)
- Lemmas too general (*loss of syntactic and semantic information*)

Solution:

- Lemmas with a subset of morphological tags

```
<f>nenahraditelná<l>nahraditelný_(*4)<t>AAFS1----1N----<r>8<g>7
                          ↓           ↓ ↓      ↓↓
                  nahraditelný_(*4)   A F      1N
                                  ⇓
                    <f>nahraditelný_(*4)<t>A*F1N</f>
                                  ⇓
                              nenahraditelná
```

Dependency Bigrams



Coocurrence Statistics

a) Contingency tables

$f(xy)$	$f(x\bar{y})$	$f(x*)$
$f(\bar{x}y)$	$f(\bar{x}\bar{y})$	$f(\bar{x}*)$
$f(*y)$	$f(*\bar{y})$	N

Example

	X=black	$X \neq$ black	X
Y=market	black market	new market	market
$Y \neq$ market	black horse	new horse	horse
Y	black	new	(all)

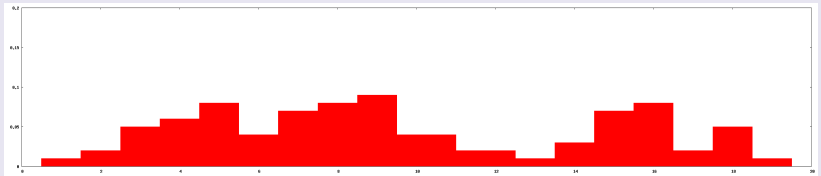
b) Contexts

C_w	global context of word w
C_{xy}	globall context of bigram xy
C^l_{xy}	left immediate context of xy
C^r_{xy}	right immediate context of xy

Example

dobrá situace . Kapitálový trh je však stále nelikvidní
že to není samostatný trh a že je součástá širšího
bariérách v přístupu na trh , cenových rozdílech ,
banky . Americký akciový trh byl za silného obchodování
jít se svou kuží na trh . Pro vydán i mluvila

Context word probability distribution $P(w_i|x)$



Types of Association Measures

1. *"Collocations are very frequent word combinations."*
   - ML estimations of joint and conditional probabilities
2. *"Collocation components occur together more often than by a chance."*
   - Mutual information and derived measures
   - Statistical tests of independence
   - Likelihood measures
   - Other heuristic association measures and coefficients
3. *"Collocations occur as units in a (inf.-theoretically) noisy environment."*
   - Immediate context measures
4. *"Collocations occur in different contexts than their components."*
   - Information-theory measures
   - Information-retrieval similarity measures

Total: 84 association measures + 3 morphosyntactic features

Data

Source: Prague Dependency Treebank v 1.0
Sentences: 81,614
Word forms: 1,255,590
Dependency bigram types: 202,171
Reference bigram types ($f>5$): 21,597
Reference collocation candidates (relevant POS): 8,904

- Data manually annotated according association strength.

4	*idioms and completely non-compositional expressions*	7
3	*partially non-compositional phrases, technical terms*	201
2	*names of persons, geographical places, and other entities*	2,698
1	*frequent compositional usages*	484
0	*non-collocations*	5,514

- All association measures computed for all bigrams.
- Comparison by precision-recall curves (no thresholds).

Data

Source: Prague Dependency Treebank v 1.0
Sentences: 81,614
Word forms: 1,255,590
Dependency bigram types: 202,171
Reference bigram types ($f>5$): 21,597
Reference collocation candidates (relevant POS): 8,904

- Data manually annotated according association strength.

4	*idioms and completely non-compositional expressions*	
3	*partially non-compositional phrases, technical terms*	2,906
2	*names of persons, geographical places, and other entities*	
1	*frequent compositional usages*	
0	*non-collocations*	5,998

- All association measures computed for all bigrams.
- Comparison by precision-recall curves (no thresholds).

Data

Source: Prague Dependency Treebank v 1.0
Sentences: 81,614
Word forms: 1,255,590
Dependency bigram types: 202,171
Reference bigram types ($f>5$): 21,597
Reference collocation candidates (relevant POS): 8,904

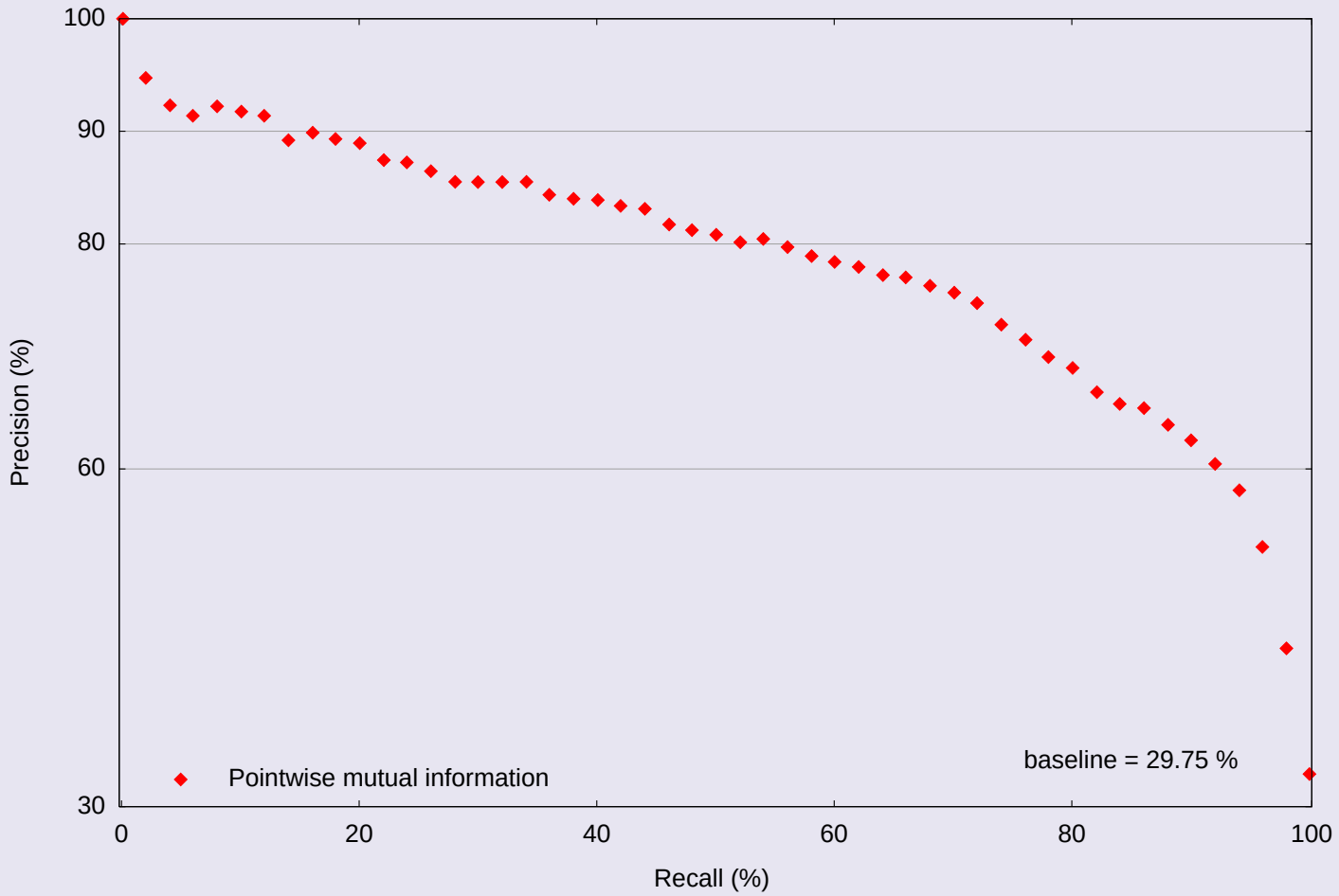
- Data manually annotated according association strength.

4	*idioms and completely non-compositional expressions*	
3	*partially non-compositional phrases, technical terms*	29 %
2	*names of persons, geographical places, and other entities*	
1	*frequent compositional usages*	
0	*non-collocations*	71 %

- All association measures computed for all bigrams.
- Comparison by precision-recall curves (no thresholds).

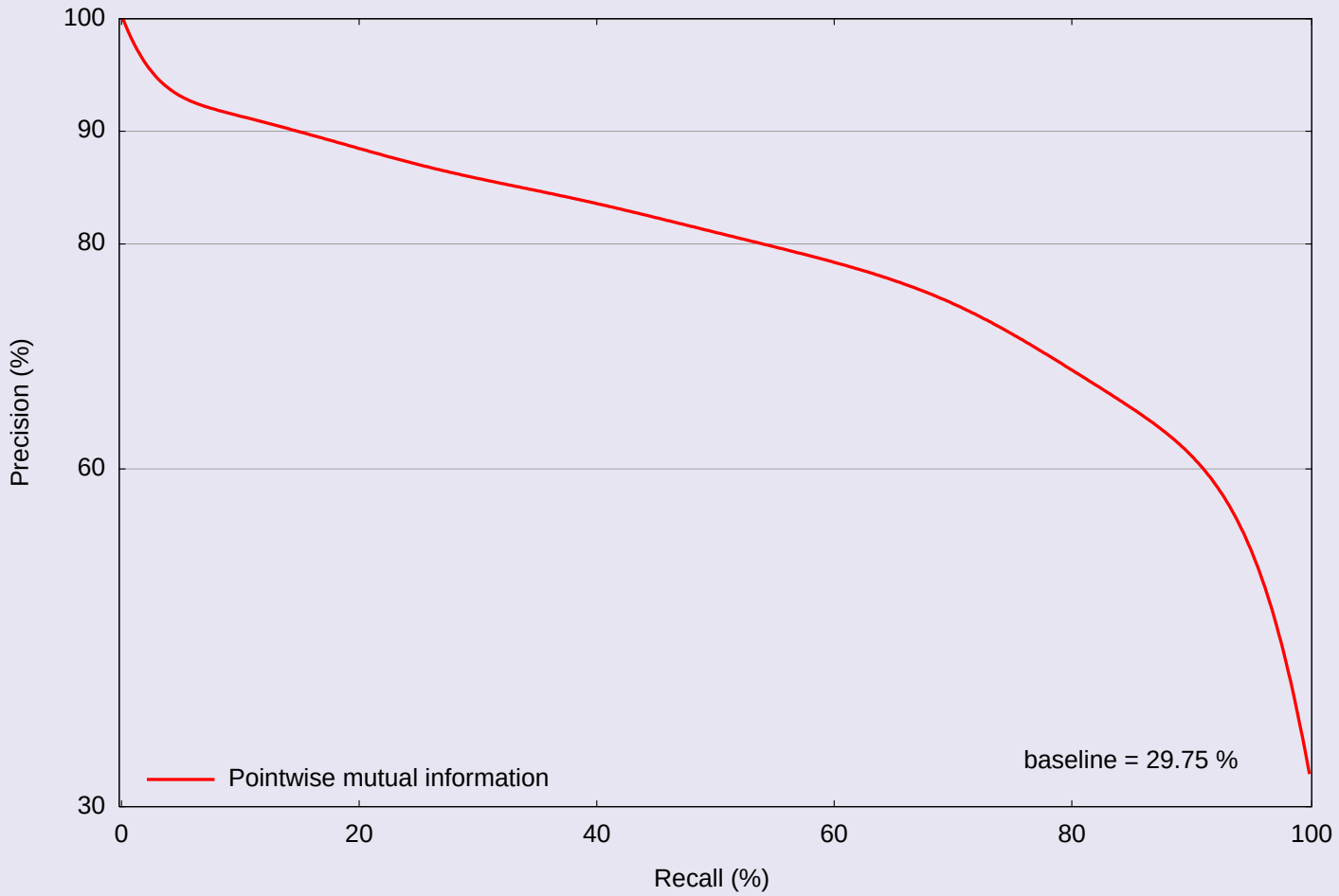
Precision-Recall Curves

Precision-Recall



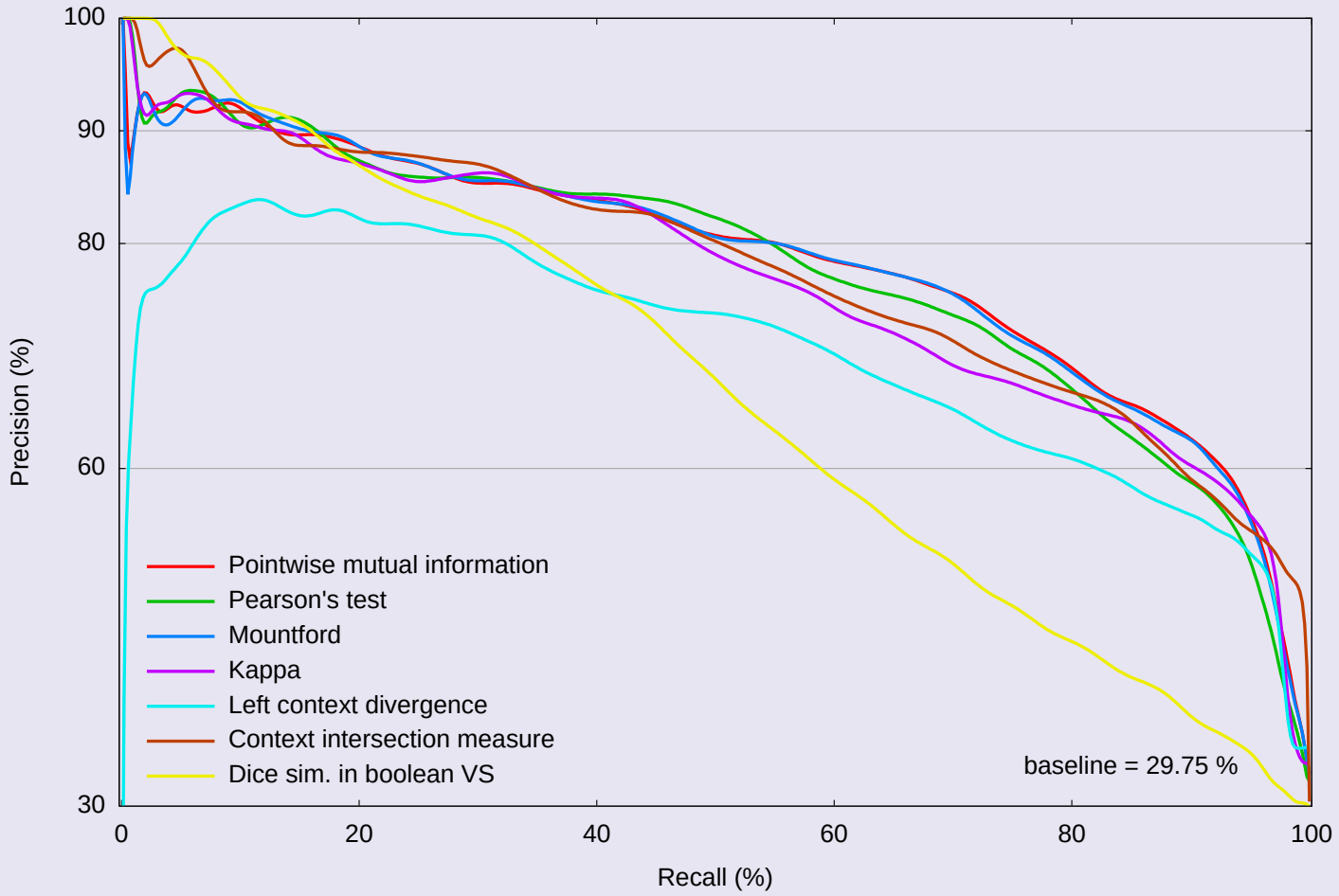
Precision-Recall Curves

Precision-Recall



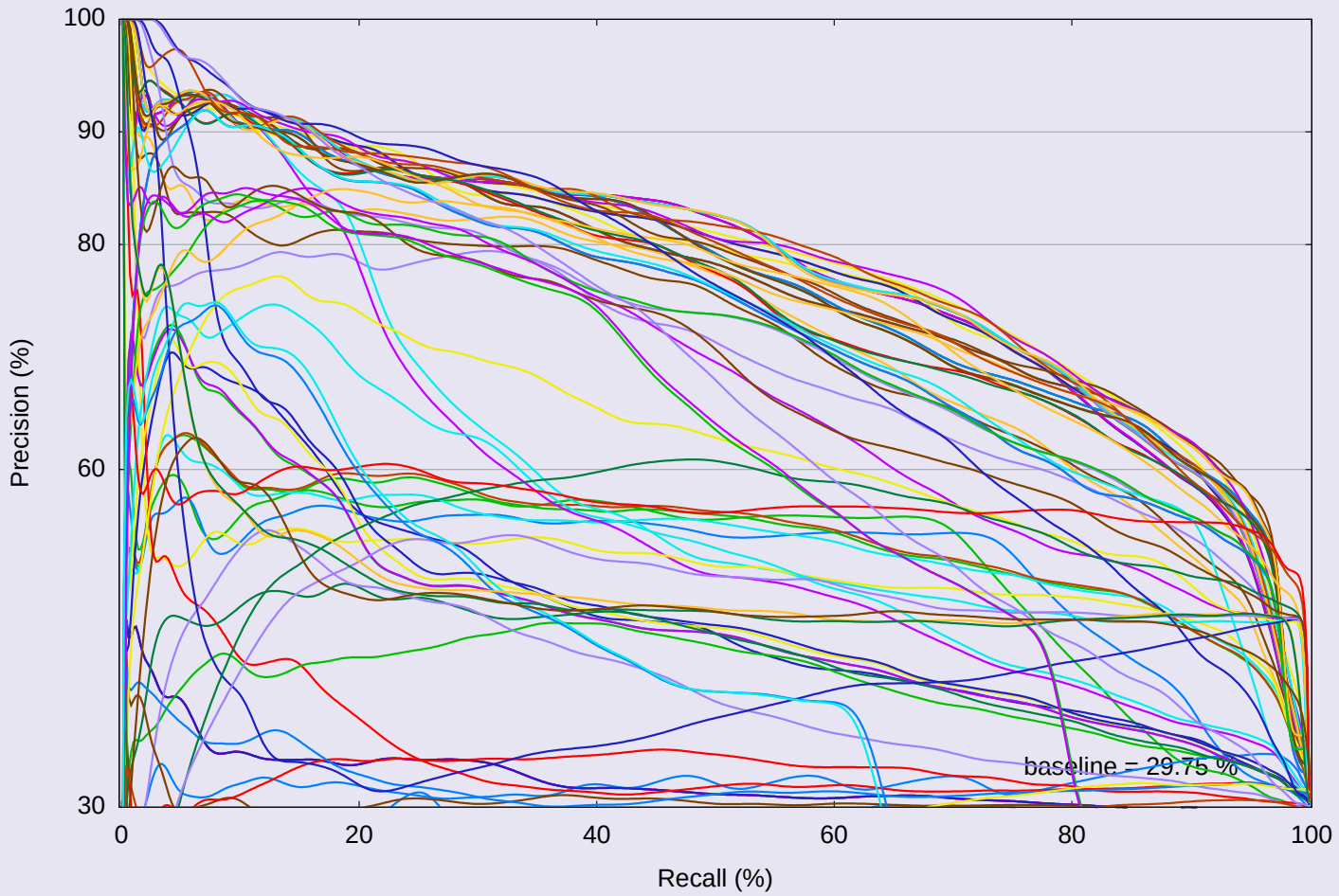
The Best Methods

Precision-Recall curves of the best association measures of each group



All Results

Precision-Recall curves of all association measures

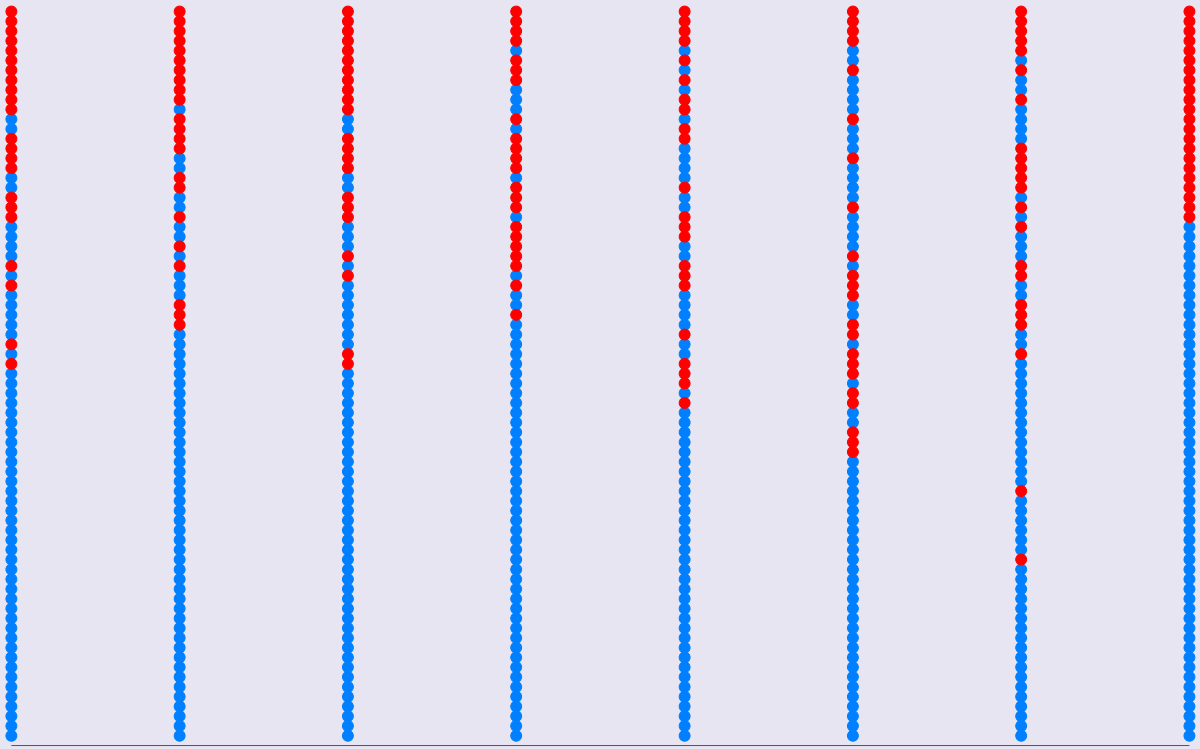


Outline

Motivation I

- Can we combine the association measures to get better results?

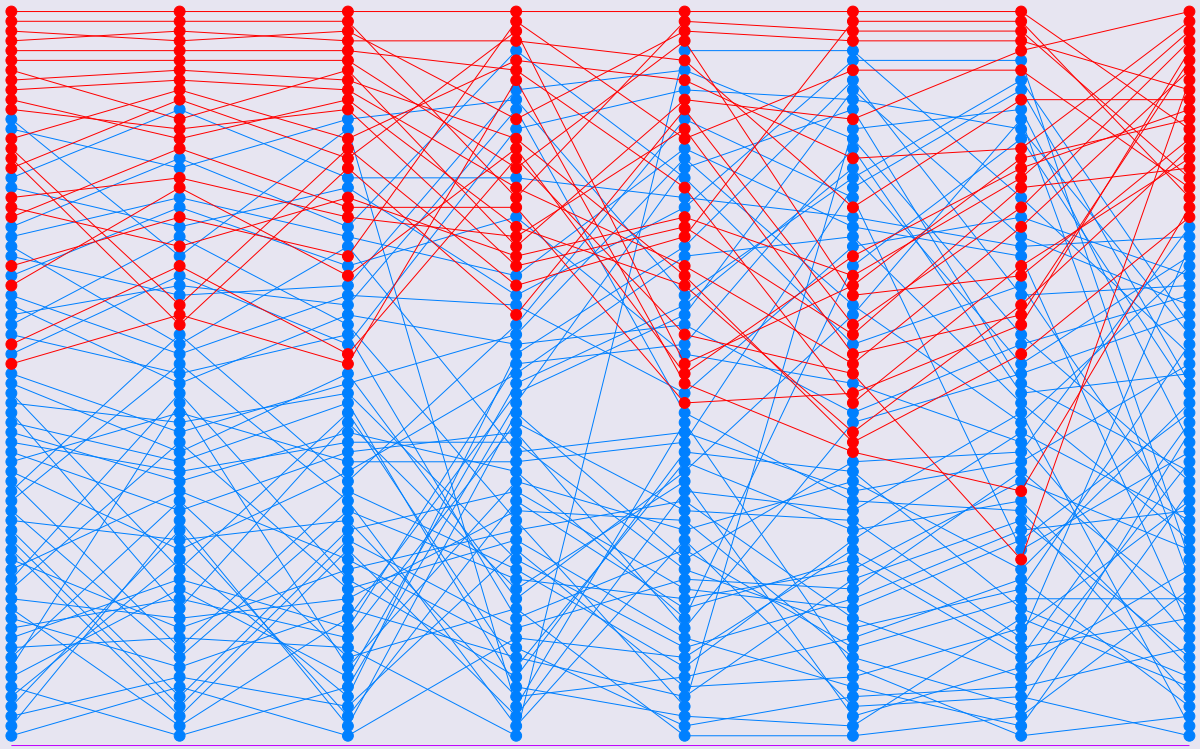
Candidates ranking by different association measures



Motivation I

- Can we combine the association measures to get better results?

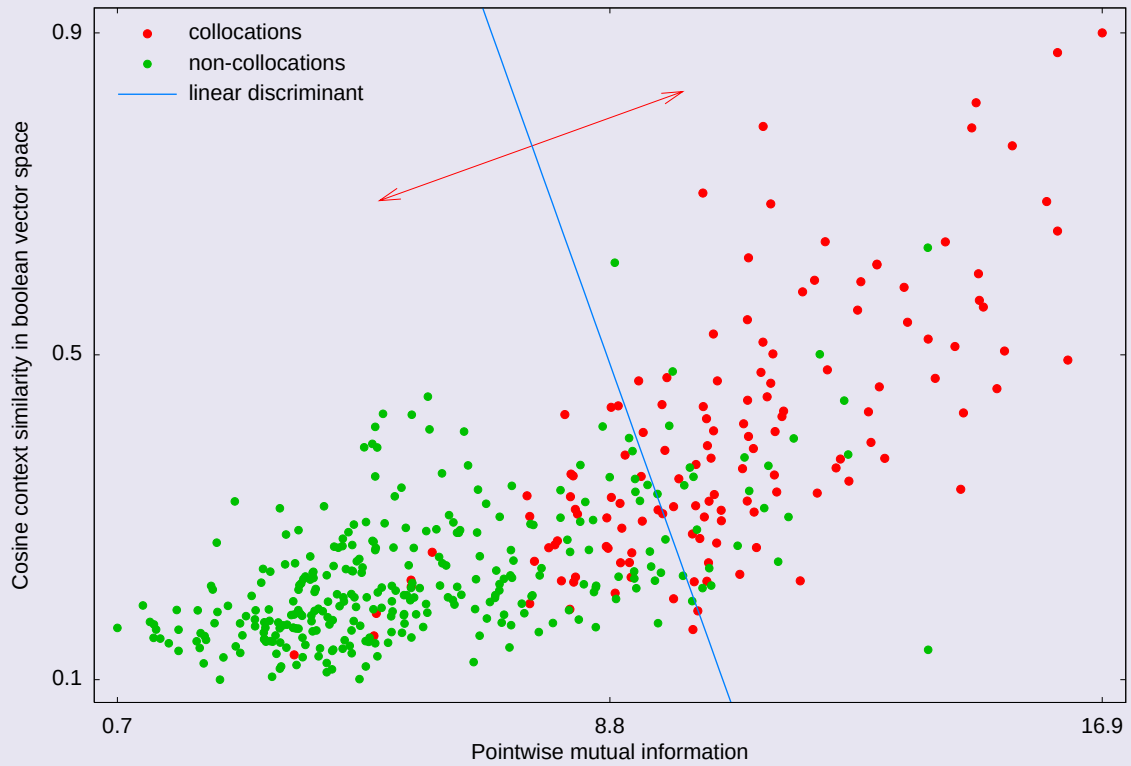
Candidates ranking by different association measures



Motivation II

- Can we combine the association measures to get better results?

Data visualization in 2D using two association measures



Combining Multiple Methods

Voting

- Each method votes whether the candidate is or is not a collocation.
- The final vote depends on the majority of the these votes.

$$\begin{array}{cccccc} x_1, & x_2, & x_3, & x_4 & \ldots & x_n \\ \downarrow & \downarrow & \downarrow & \downarrow & & \downarrow \\ 1 & 0 & 1 & 1 & \ldots & 0 \end{array} \Rightarrow y$$

Liner combination

- Each association score is weighted by its coefficient.
- The final score is defined as combination of these weighted scores.

$$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \ldots + \beta_n x_n = y$$

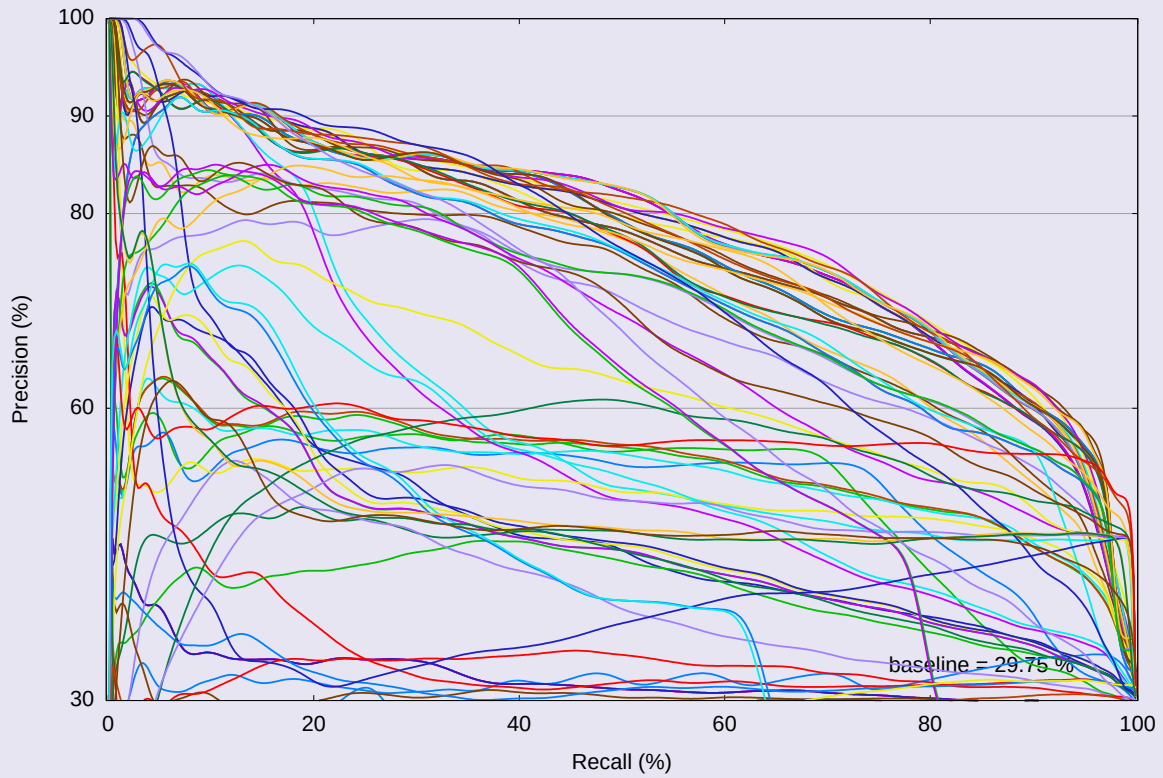
Logistic Regression

$$P(\mathbf{x}\ is\ collocation) = \frac{\exp^{\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_n x_n}}{1+\exp^{\beta_0+\beta_1 x_1 \ldots+\beta_2 x_2+\beta_n x_n}}$$

Logistic Regression

$$P(\mathbf{x}\ is\ collocation) = \frac{\exp^{\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_n x_n}}{1+\exp^{\beta_0+\beta_1 x_1 \ldots+\beta_2 x_2+\beta_n x_n}}$$

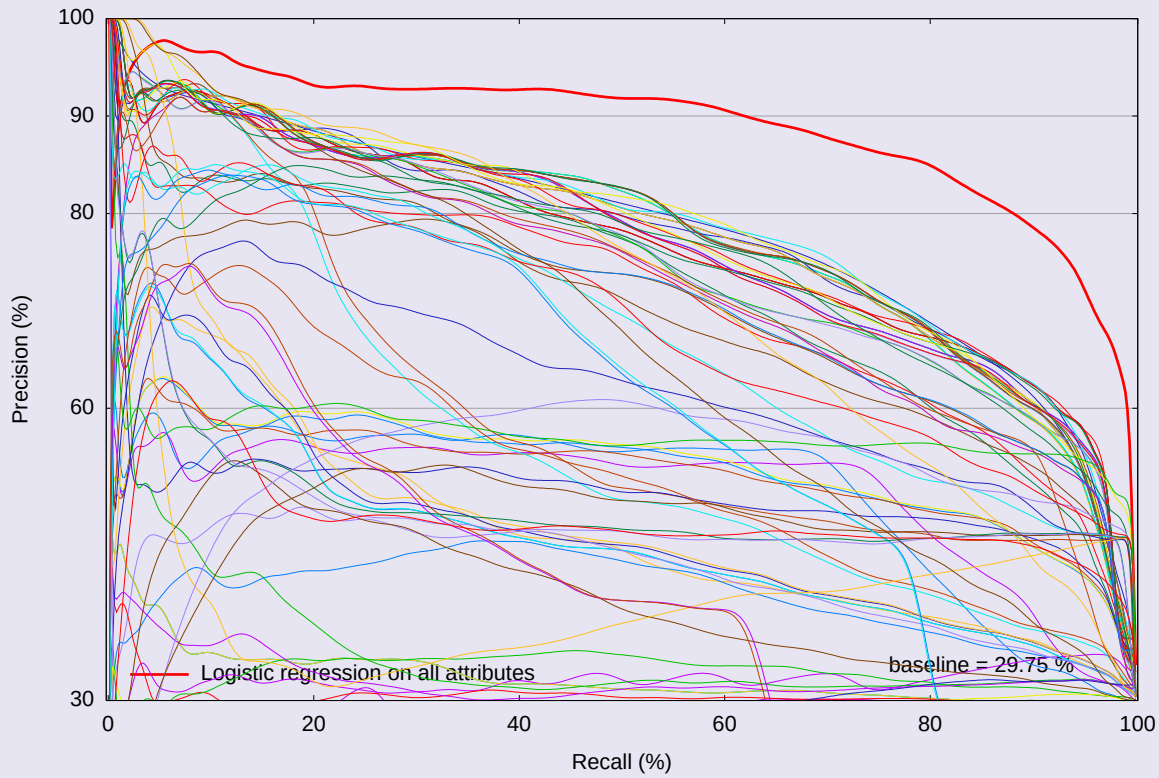
Combination of multiple methods by logistic regression



Logistic Regression

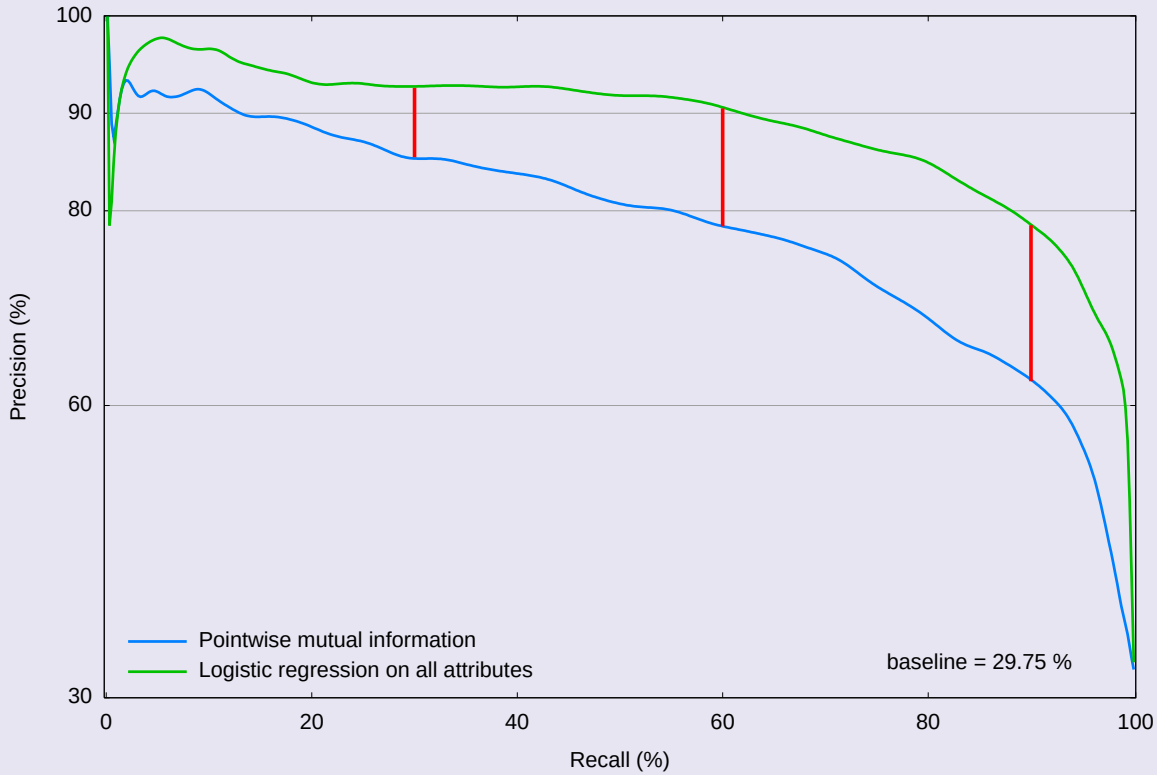
$$P(\mathbf{x}\ is\ collocation) = \frac{\exp^{\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_n x_n}}{1+\exp^{\beta_0+\beta_1 x_1 \ldots+\beta_2 x_2+\beta_n x_n}}$$

Combination of multiple methods by logistic regression



Logistic Regression: Results I

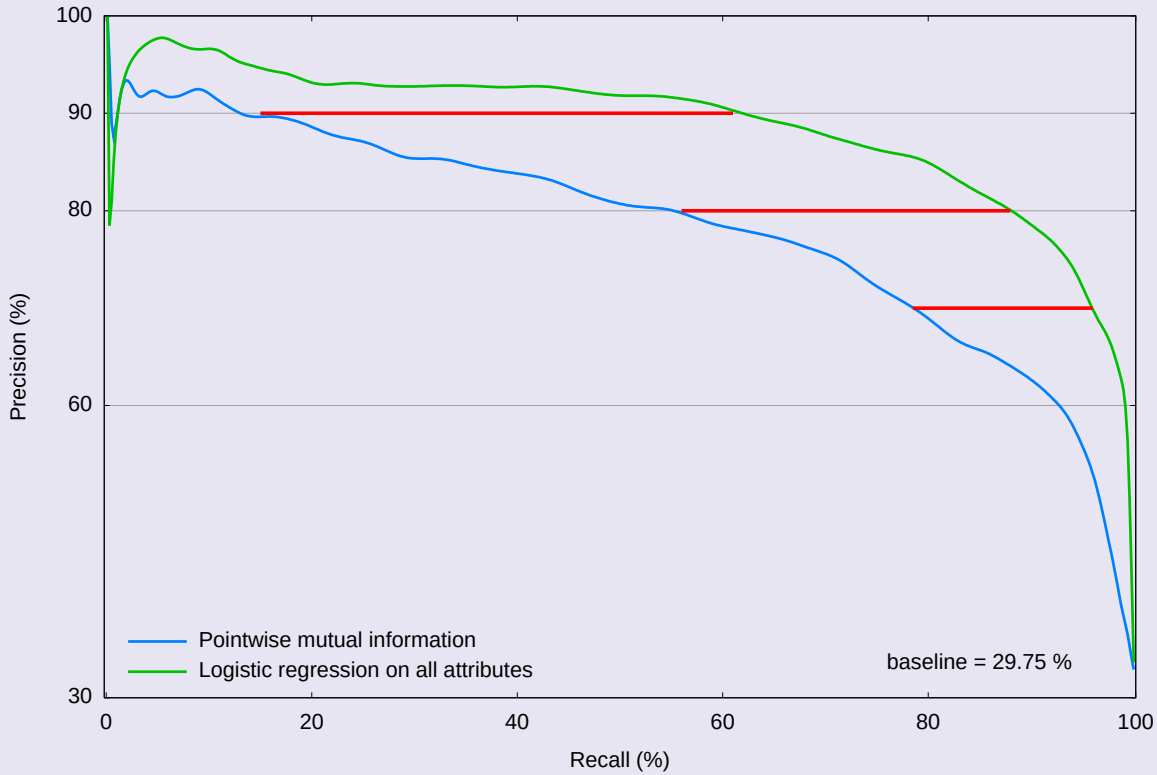
Precision improvement



Recall	30	60	90
P. mutual information	85.5	78.4	62.5
Logistic regression	92.6	89.5	84.5
Absolute improvement	7.1	11.1	**22.0**
Relative improvement	8.3	14.2	**35.2**

Logistic Regression: Results II

Recall improvement



Precision	90	80	70
P. mutual information	16.3	56.0	78.0
Logistic regression	55.8	86.7	96.7
Absolute improvement	**39.2**	30.7	17.7
Relative improvement	**242.3**	54.8	23.9

Attribute Selection

Can we reduce the number of combined association measures?

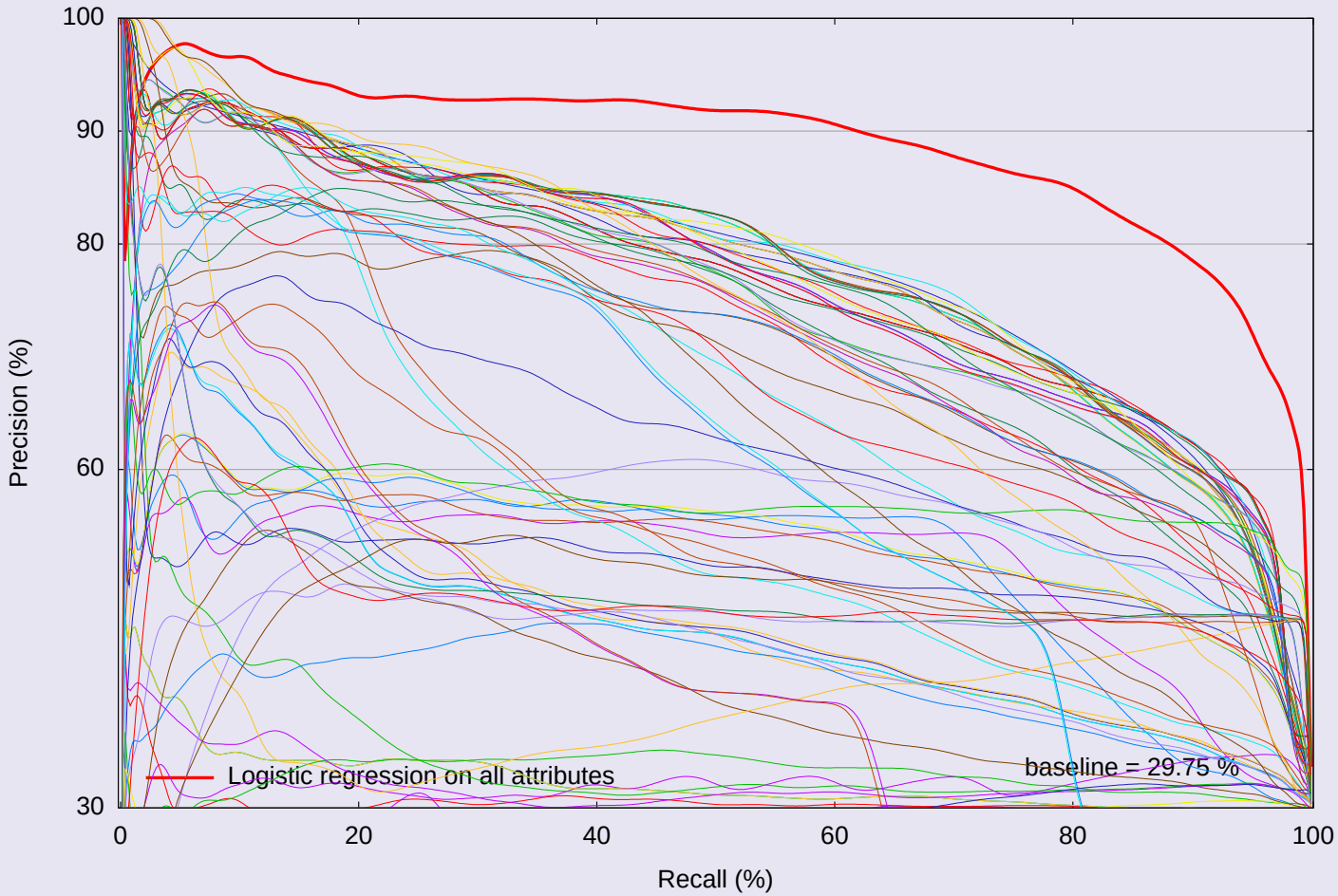
Greedy (stepwise) attribute selection:

1. Start with a full set of attributes.
2. Estimate parametres of the model.
3. Remove the attribute that minimally reduces the performance.
4. Repeat until the performance changes significantly.

Result: 87 reduced to 17

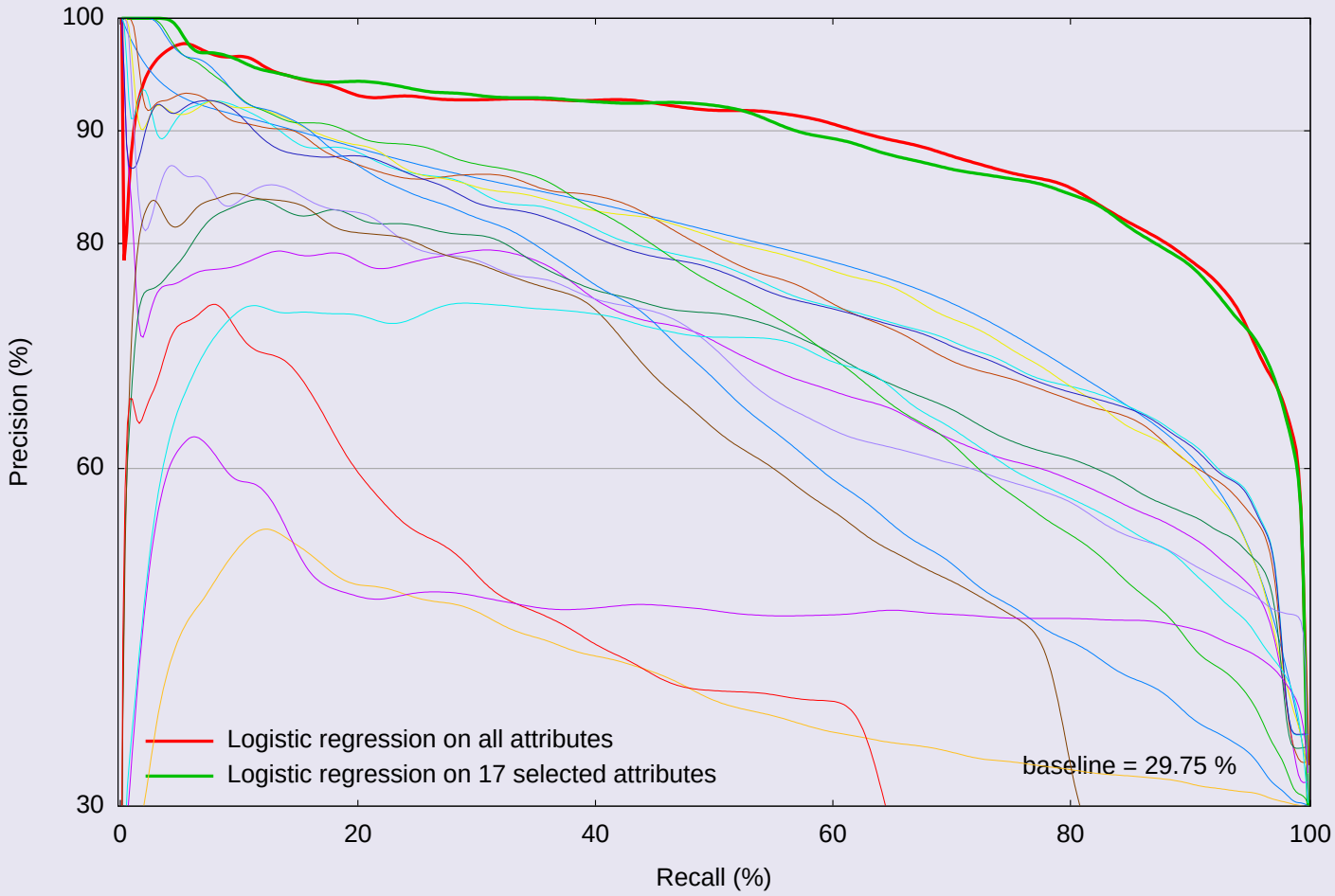
Attribute Selection: Beginning

Logistic regression on all attributes (84 +3 attributes)



Attribute Selection: End

Greedy attribute selection using logistic regression (17 attributes)



Outline

Summary

Achieved results

- Empirical evaluation of 84 association measures.
  Pointwise mutual information evaluated as one of the best measures.
- Statistical combination of multiple association measures.
  Linear logistic regression gives significant performance improvement.
- Selection of the best subset of association measures.
  Greedy algorithm reduced number of association measures to 17.

Outlook

- Multiple annotation of the reference data.
- Employing other classification (ranking) methods.

That's all folks . . .

Thank you!

Association Measures I

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**	$\frac{\hat{f}(xy) - f(xy)\log\hat{f}(xy) + \log f(xy)!}{\log N}$

Association Measures II

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**	$\frac{a}{\sqrt{(a+b)(a+c)}}$

Association Measures III

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**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$

Association Measures IV

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**	$-\sum_w P(w\|C^r_{xy}) \log P(w\|C^r_{xy})$

Association Measures V

*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**	$\sum_w P(w\|C_y)\log\frac{P(w\|C_y)}{P(w\|C_x)}$

Association Measures VI

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*·*idf* vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x:w_i\in C_x\}\|$
	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*·*idf* vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x:w_i\in C_x\}\|$
⋆85.	**Part of speech**	{*Adjective:Noun, Noun:Noun, Noun:Verb, ...*}
⋆86.	**Dependency type**	{*Attribute, Object, Subject, ...*}
87.	**Dependency structure**	{↗, ↖}