Collocation Extraction

The Statistical Approach

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University

April 12, 2005

Outline

Outline

Valence a kolokabilita

- Paradigma – třída funkčně ekvivalentních (lexikálních) jednotek.
- Virtuální paradigma – ekvivalence na základě téže vlastnosti, *vyjádřená intenzionálně*.
- Kolokační paradigma - ekvivalence na základě užší substituovatelnosti v tomtéž místě textu, vůči jinému slovu, *vyjádřená extenzionálně*.

Čermák (Syntagmatika a paradigmatika českého slova I, 1982)

"Kolokabilita je obecná schopnost slova se spojovat v textu s jinými. Pokud je vyjádřena intenzionálně mluvíme o valenci. V ostatních případech (vymezená extenzionálně) postačí širší termín kolokabilita."

"V textu realizovaná kolokabilita (alesoň) dvou paradigmat, resp. jejich prvků dává slovní spojení, kolokaci."

Kolokace

Encyklopedický slovník češtiny (2002)

"Kolokace je kombinace jazykových prvků lexikální povahy (spojení slov)."

Typy kolokací:

1. systémové
   a) pravidelné - *víceslovné termíny (kyselina sírová, cestovní kancelář)*
   b) nepravidelné - *frazémy a idiomy (ležet ladem, černá díra)*
2. textové
   a) pravidelné - *běžné kolokace užívané preferenčně (letní dovolená)*
   b) nepravidelné - *individuální autorské metafory (třeskutě vtipný)*

V některých pojetích: *kolokace = kolokace typické (běžné)*
Ve frazeologii a idiomatice: *kolokace = kolokační frazém*

Sousloví

Šmilauer (Nauka o českém jazyce, 1979)

"Sousloví je ustálené pojmenování, které vzniklo ze spojení dvou či více slov, má však význam slova jednoho a vstupuje do věty jako hotový celek."

Různé stupně sousloví:

1. nerozložitelná, slova jinak neužívaná *(křížem krážem)*
2. těsně spojené členy, neodělují se čárkou *(ve dne v noci)*
3. rozložitelná, ale jako celek metaforická *(paví oko)*
4. rozložitelná, jeden člen metaforický *(vlčí mák)*
5. oba členy významové *(jízdní rychlost)*

Motivace

Čermák (Manuál Lexikografie, 1995)

"Víceslovných lexémů (kolokací) je v češtině více než lexémů jednoslovných a pro svá specifika ve frazeologii, idiomatice a terminologii vyžadují konstituování samostatné paralelní části v rámci lexikální databáze."

Čermák (Syntagmatika a paradigmatika českého slova I, 1982)

"Jsou však případy, kdy se slovo může uzuálně spojovat s jedním, popř. několika málo dalšími a pak je nejen užitečné, ale dokonce i nutné je znát, jinak ho nemůžeme používat. Zároveň se pak dá jedině takto (extenzionálně) vystihnout jeho význam."

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."

Definitions II

Radev (1998)

"A collocation is a group of words that occur together more often than by a chance."

Manning (1999)

"A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things."

Definitions III

Bartsch (2004)

"Collocations can be defined as frequently reccurrent, relatively fixed syntagmatic combinations of two or more words."

Evert (2004)

"A collocation is a word combination whose semantic and/or syntactic properties cannot be fully predicted from those of its components, and which therefore has to be listed in a lexicon."

Characteristic Features

- **Non-compositionality**

  The meaning of a collocation is not a straightforward composition of the meaning of its parts.
  (kick the bucket, carriage return, white man)

- **Non-substitutability**

  Components of collocation cannot be substituted with a related word or a synonym.
  (yellow wine, hit the bucket, make homework)

- **Non-modifiability**

  Collocations cannot be modified or syntactically transformed.
  (give a big hand, poor as a church mice)

Other Issues

- **Translatability**

  Word by word translation to other languages is for some collocations not possible.
  (ice cream, to be right)

- **Domain specificity**

  Some expressions are collocations in their domain however in other domain can be completely compositional.
  (carriage return)

- **Subjectivity**

  Two individuals can perceive a word expression as a collocation differently.
  (game over, new company)

Outline

Some well-known problems I

- Lexicography

  Which multiword expressions to include into a lexicon?
  My new computer is a laptop computer.

- Machine translation

  Where to break a sentence into chunks?
  She likes ice cream pancakes.

- Information retrieval

  Which multiword terms to index?
  Our new friend is from New York.

- Word sense disambiguation

  How to distinguish between possible word senses?
  My uncle owns a wine yard.

Some well-known problems II

- Spell/grammar/style-checking

  Is this text written correctly?
  Meals will be served outside, weather allowing.

- Text classification and summarization

  What is this text about?
  Carriage return is necessary at the end of line.

- Language modeling (text/speech synthesis)

  How to create a fluent sentence?
  Could you hand me pepper and salt?

- Corpus-based language teaching/learning

  What kinds of multiword expressions to teach?
  He finally kicked the bucket.

Outline

The task and the Solution

To build a collocation lexicon.

Possible approaches:

1. Manual processing of a corpus and annotation of collocations.
   - \+ Good results
   - – Subjective results
   - – Slow and expensive process
2. Automatic processing of a corpus and extraction of collocations.
   - \+ Fast process
   - \+ Stable results
   - \+ Customizable output
   - – Certain error
3. Combination of the two previous.

The task and the Solution

To build a collocation lexicon.

Possible approaches:

1. Manual processing of a corpus and annotation of collocations.
   - \+ Good results
   - – Subjective results
   - – Slow and expensive process
2. Automatic processing of a corpus and extraction of collocations.
   - \+ Fast process
   - \+ Stable results
   - \+ Customizable output
   - – Certain error $\longrightarrow$ Predictable!
3. Combination of the two previous.

Outline

Collocation Extraction Methodology

- Most methods are based on verification of typical collocation properties.
- These properties are formally described by mathematical formulas that determine degree of association between components of collocation.
- 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

červený kříž	*15.66*
řádová čárka	*14.01*
aritmetická operace	*10.52*
podavač papíru	*10.17*
systém typu	*3.54*
na další	*0.54*
program v	*0.35*
úroveň být	*0.25*

Classification

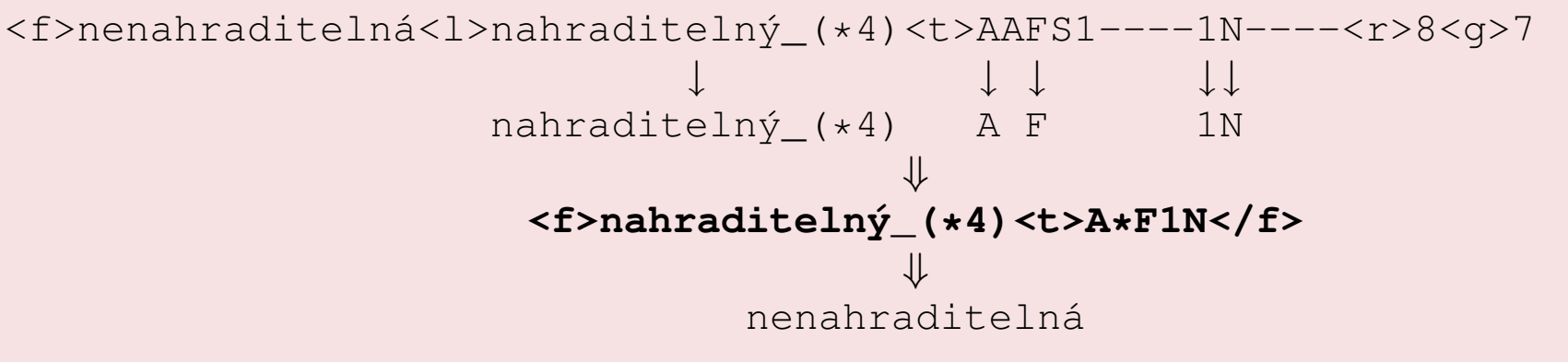
červený kříž	*1*
řádová čárka	*1*
aritmetická operace	*1*
podavač papíru	*1*
systém typu	*0*
na další	*0*
program v	*0*
úroveň být	*0*

Candidate Extraction Process

1. Extracting all possible candidates for collocations
   - Consequent word n-grams
   - Sliding window
   - Syntactic structures
2. Collecting cooccurrence statistics
   - Frequency of word and n-gram occurrences
   - Immediate contexts
   - Average contexts

Word Base Forms

- Full word forms too specific (*rich morphology*)
- Lemmas too general (*loss of semantic information*)
- Solution: lemmas with a subset of morphological tags



Morphological and Syntactical Patterns

Useful for:

- Filtering
- Classification
- Distinction

Part-of-speech

A N	*lineární funkce*
N N	*následník trůnu*
D A N	*objektově orientovaný jazyk*
N A N	*zbraně hromadného ničení*
V R N	*přijít k sobě*

Dependency type

Atr	*cenný papír*
Sb	*ceny stoupají*
Obj	*dát přednost*
Adv	*zdravotně postižený*

Dependency direction

velký výr	*výr velký*

Outline

Association Measures

Basic approaches:

1. Estimations of bigram and unigram probabilities
2. Mutual information and derived measures
3. Statistical tests of independence
4. Likelihood measures
5. Other heuristic association measures and coefficients
6. Immediate context measures
7. Information-theory average context measures
8. Average context similarity measures

Occurrence Probability

Model: Words are generated randomly and independently.
Hypothesis: Collocations are frequent word combinations.

Example

```
... ... doba .. ... ... ...
... doba ledová ... ... ...
ledová ... ... ... doba ...
... ... doba ledová ... ...
... doba .. ... ... ... ...
... ... doba ledová ... ...
... ... ... ... ledová ....
```

$$c(\textit{doba ledova}) = 3$$
$$N = 1000$$

$$P(\textit{doba ledova}) = \frac{c(\textit{doba ledova})}{N} = \frac{3}{1000}$$

Pointwise Mutual Information

Model: Words are generated randomly and independently.
Hypothesis: Mutual information of collocation component occurrences is high.

Example

```
... ... doba .. ... ... ...
... doba ledová ... ... ...
ledová ... ... ... doba ...
... ... doba ledová ... ...
... doba .. ... ... ... ...
... ... doba ledová ... ...
... ... ... ... ledová ....
```

$$P(doba) = \frac{6}{1000} = 0.006$$

$$P(ledova) = \frac{5}{1000} = 0.005$$

$$P(doba\,ledova) = \frac{3}{1000} = 0.003$$

$$I(doba\,ledova) = log\frac{P(doba\,ledova)}{P(doba) \times P(ledova)}$$

$$= log\frac{0.003}{0.005 \times 0.006}$$

t test

Model: Words are generated randomly and independently.
Hypothesis: Collocations occur together more often than by a chance.

Example

```
... ... doba .. ... ... ...
... doba ledová ... ... ...
ledová ... ... ... doba ...
... ... doba ledová ... ...
... doba .. ... ... ... ...
... ... doba ledová ... ...
... ... ... ... ledová ....
```

$$P(doba) = \frac{6}{1000} = 0.006$$

$$P(ledova) = \frac{5}{1000} = 0.005$$

$$P(doba\,ledova) = \frac{3}{1000} = 0.003$$

$$t = \frac{P(doba\,ledova) - P(doba) \times P(ledova)}{\sqrt{P(doba) \times P(ledova)/N}}$$

$$= \frac{0.003 - 0.005 \times 0.006}{\sqrt{0.005 \times 0.006/1000}}$$

Contingency Table Measures

Model: Words are generated randomly and independently.
Hypothesis: Occurrence of collocation components is not independent.

bigram: *xy*

	$X{=}x$	$X{\neq}x$
$Y{=}y$	a	b
$Y{\neq}y$	c	d

Example

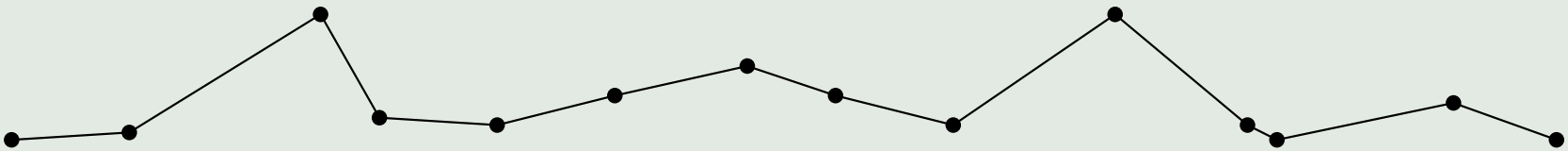
	$X{=}$černý	$X{\neq}$černý
$Y{=}$trh	černý trh	domácí trh
$Y{\neq}$trh	černý čaj	zelená tráva

$$\frac{a}{a+b+c+d} \qquad \frac{b+c}{d+e} \qquad \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}} \qquad \frac{a}{max(a+b,a+c)}$$

Immediate Context Analysis

Model: Words are generated randomly depending on the preceding words.
Hypothesis: Collocations occur as units in a noisy environment.

Example

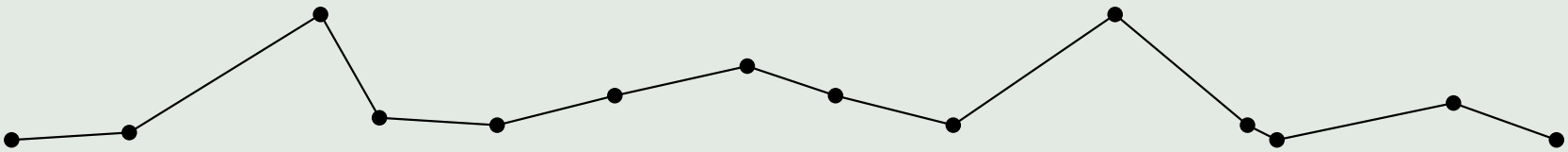


Český kapitálový trh dnes ovlivňil pokles cen všech cenných papírů a zejména akcií.

Immediate Context Analysis

Model: Words are generated randomly depending on the preceding words.
Hypothesis: Collocations occur as units in a noisy environment.

Example



Český *kapitálový trh* dnes ovlivňil pokles cen všech *cenných papírů* a zejména akcií.

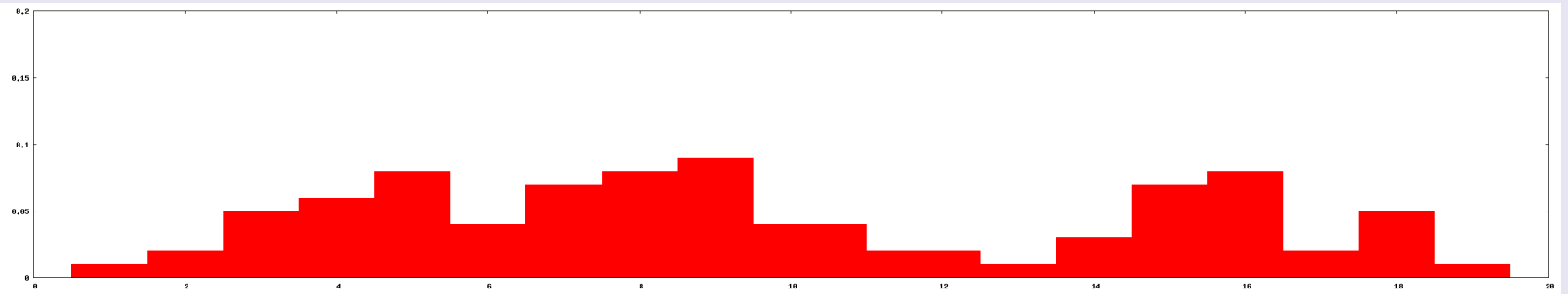
Average Context Similarity

Model: Each word is determined by its context.
Hypothesis: Average contexts of collocation components are different and differ also from the average context of the collocation.

Example

zlepšení 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 zejména

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



Association Measures I

1.	**Dependency structure**	{0:1,2:0}
⋆2.	**Part of speech**	{A:N,N:N,N:V,V:N,R:N,D:V,...}
⋆3.	**Dependency type**	{Atr:Head,Head:Atr,Head:Obj,...}
4.	**Number of components**	2
5.	**Mean component offset**	$\frac{1}{n}\sum_{i=1}^{n} d_i$
6.	**Variance component offset**	$\frac{1}{n-1}\sum_{i=1}^{n} (d_i - \bar{d})^2$
7.	**Joint probability**	$P(xy)$
8.	**Conditional probability**	$P(y\|x)$
9.	**Reverse conditional prob.**	$P(x\|y)$
⋆10.	**Pointwise mutual information**	$\log \frac{P(xy)}{P(x*)P(*y)}$
11.	**Mutual dependency (*MD*)**	$\log \frac{P(xy)^2}{P(x*)P(*y)}$
12.	**Log frequency biased *MD***	$\log \frac{P(xy)^2}{P(x*)P(*y)} + \log P(xy)$
13.	**Normalized expectation**	$\frac{2f(xy)}{f(x*)+f(*y)}$
⋆14.	**Mutual expectation**	$\frac{2f(xy)}{f(x*)+f(*y)} \cdot P(xy)$
15.	**Salience**	$\log \frac{P(xy)^2}{P(x*)P(*y)} \cdot \log f(xy)$
16.	**Pearson's χ^2 test**	$\sum_{ij} \frac{(f_{ij} - \hat{f}_{ij})^2}{\hat{f}_{ij}}$
17.	**Fisher's exact test**	$\frac{f(x*)!f(\bar{x}*)!f(*y)!f(*\bar{y})!}{N!f(xy)!f(x\bar{y})!f(\bar{x}y)!f(\bar{x}\bar{y})!}$

Association Measures II

18.	**t test**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{f(xy)(1-(f(xy)/N))}}$
19.	**z score**	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(\hat{f}(xy)/N))}}$
20.	**Poison significance measure**	$\frac{\hat{f}(xy)-f(xy)\log\hat{f}(xy)+\log f(xy)!}{\log N}$
21.	**Log likelihood ratio**	$-2\sum_{ij} f_{ij}\log\frac{f_{ij}}{\hat{f}_{ij}}$
22.	**Squared log likelihood ratio**	$-2\sum_{ij}\frac{\log f_{ij}^2}{\hat{f}_{ij}}$
Association coefficients:		
23.	**Russel–Rao**	$\frac{a}{a+b+c+d}$
24.	**Sokal-Michiner**	$\frac{a+d}{a+b+c+d}$
⋆25.	**Rogers–Tanimoto**	$\frac{a+d}{a+2b+2c+d}$
26.	**Hamann**	$\frac{(a+d)-(b+c)}{a+b+c+d}$
27.	**Third Sokal–Sneath**	$\frac{b+c}{a+d}$
28.	**Jaccard**	$\frac{a}{a+b+c}$
⋆29.	**First Kulczynsky**	$\frac{a}{b+c}$
30.	**Second Sokal–Sneath**	$\frac{a}{a+2(b+c)}$
31.	**Second Kulczynski**	$\frac{1}{2}(\frac{a}{a+b}+\frac{a}{a+c})$
32.	**Fourth Sokal–Sneath**	$\frac{1}{4}(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c})$

Association Measures III

33.	**Odds ratio**	$\frac{ad}{bc}$
34.	**Yulle's** ω	$\frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$
⋆35.	**Yulle's** Q	$\frac{ad-bc}{ad+bc}$
36.	**Driver–Kroeber**	$\frac{a}{\sqrt{(a+b)(a+c)}}$
37.	**Fifth Sokal–Sneath**	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
38.	**Pearson**	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
39.	**Baroni–Urbani**	$\frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$
40.	**Braun–Blanquet**	$\frac{a}{\max(a+b,a+c)}$
41.	**Simpson**	$\frac{a}{\min(a+b,a+c)}$
42.	**Michael**	$\frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$
43.	**Mountford**	$\frac{2a}{2bc+ab+ac}$
44.	**Fager**	$\frac{a}{\sqrt{(a+b)(a+c)}}-\frac{1}{2}\max(b,c)$
45.	**Unigram subtuples**	$\log\frac{ad}{bc}-3.29\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$
46.	***U* cost**	$\log(1+\frac{\min(b,c)+a}{\max(b,c)+a})$
47.	***S* cost**	$\log(1+\frac{\min(b,c)}{a+1})^{-\frac{1}{2}}$
48.	***R* cost**	$\log(1+\frac{a}{a+b})\cdot\log(1+\frac{a}{a+c})$
49.	***T* combined cost**	$\sqrt{U\times S\times R}$

Association Measures IV

50.	**Phi**	$\frac{P(xy)-P(x*)P(*y)}{\sqrt{P(x*)P(*y)(1-P(x*))(1-P(*y))}}$
51.	**Kappa**	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
52.	***J* measure**	$\max[P(xy)\log\frac{P(y\|x)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{y}\|x)}{P(*\bar{y})},$ $P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
53.	**Gini index**	$\max[P(x*)(P(y\|x)^2+P(\bar{y}\|x)^2)-P(*y)^2$ $+P(\bar{x}*)(P(y\|\bar{x})^2+P(\bar{y}\|\bar{x})^2)-P(*\bar{y})^2,$ $P(*y)(P(x\|y)^2+P(\bar{x}\|y)^2)-P(x*)^2$ $+P(*\bar{y})(P(x\|\bar{y})^2+P(\bar{x}\|\bar{y})^2)-P(\bar{x}*)^2]$
54.	**Confidence**	$\max[P(y\|x),P(x\|y)]$
55.	**Laplace**	$\max[\frac{NP(xy)+1}{NP(x*)+2},\frac{NP(xy)+1}{NP(*y)+2}]$
56.	**Conviction**	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})},\frac{P(\bar{x}*)P(*y)}{P(\bar{x}y)}]$
57.	**Piatersky–Shapiro**	$P(xy)-P(x*)P(*y)$
58.	**Certainity factor**	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)},\frac{P(x\|y)-P(x*)}{1-P(x*)}]$
59.	**Added value (*AV*)**	$\max[P(y\|x)-P(*y),P(x\|y)-P(x*)]$
⋆60.	**Collective strength**	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)}\cdot$ $\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
61.	**Klosgen**	$\sqrt{P(xy)}\cdot AV$

Association Measures V

Context measures:		
⋆ 62.	**Context entropy**	$-\sum_w P(w\|C_{xy}) \log P(w\|C_{xy})$
63.	**Left context entropy**	$-\sum_w P(w\|C^l_{xy}) \log P(w\|C^l_{xy})$
64.	**Right context entropy**	$-\sum_w P(w\|C^r_{xy}) \log P(w\|C^r_{xy})$
⋆ 65.	**Left context divergence**	$P(x*) \log P(x*)$ $-\sum_w P(w\|C^l_{xy}) \log P(w\|C^l_{xy})$
66.	**Right context divergence**	$P(*y) \log P(*y)$ $-\sum_w P(w\|C^r_{xy}) \log P(w\|C^r_{xy})$
67.	**Cross entropy**	$-\sum_w P(w\|C_x) \log P(w\|C_y)$
68.	**Reverse cross entropy**	$-\sum_w P(w\|C_y) \log P(w\|C_x)$
69.	**Intersection measure**	$\frac{2\|C_x \cap C_y\|}{\|C_x\|+\|C_y\|}$
70.	**Euclidean norm**	$\sqrt{\sum_w (P(w\|C_x) - P(w\|C_y))^2}$
71.	**Cosine norm**	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2 \cdot \sum_w P(w\|C_y)^2}$
72.	***L1* norm**	$\sum_w \|P(w\|C_x) - P(w\|C_y)\|$
73.	**Confusion probability**	$\sum_w \frac{P(x\|C_w)P(y\|C_w)P(w)}{P(x*)}$
74.	**Reverse confusion prob.**	$\sum_w \frac{P(y\|C_w)P(x\|C_w)P(w)}{P(*y)}$
⋆ 75.	**Jensen-Shannon diverg.**	$\frac{1}{2}[D(p(w\|C_x)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))$ $+D(p(w\|C_y)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))]$

Association Measures VI

76.	**Cosine of pointwise *MI***	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2}\cdot\sqrt{\sum_w MI(w,y)^2}}$
⋆77.	**KL divergence**	$\sum_w P(w\|C_x)\log\frac{P(w\|C_x)}{P(w\|C_y)}$
⋆78.	**Reverse KL divergence**	$\sum_w P(w\|C_y)\log\frac{P(w\|C_y)}{P(w\|C_x)}$
79.	**Skew divergence**	$D(p(w\|C_x)\|\|\alpha p(w\|C_y)+(1-\alpha)p(w\|C_x))$
80.	**Reverse skew divergence**	$D(p(w\|C_y)\|\|\alpha p(w\|C_x)+(1-\alpha)p(w\|C_y))$
81.	**Phrase word coocurrence**	$\frac{1}{2}(\frac{f(x\|C_{xy})}{f(xy)}+\frac{f(y\|C_{xy})}{f(xy)})$
82.	**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:	$\frac{1}{2}(\cos(\mathbf{c}_x,\mathbf{c}_{xy})+\cos(\mathbf{c}_y,\mathbf{c}_{xy}))$
		$\mathbf{c}_z=(z_i);\ \cos(\mathbf{c}_x,\mathbf{c}_y)=\frac{\sum x_i y_i}{\sqrt{\sum {x_i}^2}\cdot\sqrt{\sum {y_i}^2}}$
⋆83.	**in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
84.	**in *tf* vector space**	$z_i=f(w_i\|C_z)$
85.	**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:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$
		$\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_i y_i}{\sum {x_i}^2+\sum {y_i}^2}$
⋆86.	**in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
⋆87.	**in *tf* vector space**	$z_i=f(w_i\|C_z)$
⋆88.	**in *tf*·*idf* vector space**	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x:w_i\in C_x\}\|$

Outline

Data

- Source: Prague Dependency Treebank v 1.0
- Sentences: 81 614
- Word forms: 1 255 590
- Dependency bigram types: 202 171
- Reference bigram types: 21 597
- Reference collocation candidates: 8 904

Reference data (collocation candidates) manually evaluated and annotated according association strength.

4	*kámen úrazu, železná opona*	7
3	*černý trh, pata kolmice*	201
2	*atomová energie, tlustá kniha*	2 698
1	*na Slovensko, do Portugalska*	484
0	*(non-collocations)*	5 514

Data

- Source: Prague Dependency Treebank v 1.0
- Sentences: 81 614
- Word forms: 1 255 590
- Dependency bigram types: 202 171
- Reference bigram types: 21 597
- Reference collocation candidates: 8 904

Reference data (collocation candidates) manually evaluated and annotated according association strength.

4 3 2	*kámen úrazu, železná opona* *černý trh, pata kolmice* *atomová energie, tlustá kniha*	2 906
1 0	*na Slovensko, do Portugalska* *(non-collocations)*	5 998

Data

- Source: Prague Dependency Treebank v 1.0
- Sentences: 81 614
- Word forms: 1 255 590
- Dependency bigram types: 202 171
- Reference bigram types: 21 597
- Reference collocation candidates: 8 904

Reference data (collocation candidates) manually evaluated and annotated according association strength.

4 3 2	*kámen úrazu, železná opona* *černý trh, pata kolmice* *atomová energie, tlustá kniha*	30 %
1 0	*na Slovensko, do Portugalska* *(non-collocations)*	70 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	*15.66*
železná opona	*15.23*
řádová čárka	*14.01*
kupónová knížka	*13.83*
autor knihy	*11.05*
aritmetická operace	*10.52*
podavač papíru	*10.17*
nová kniha	*10.09*
kulatý stůl	*7.03*
nová vlna	*6.59*
čerpací stanice	*6.04*
systém typu	*3.54*
centrum města	*1.54*
na další	*0.54*
program v	*0.35*
úroveň být	*0.25*

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	0
aritmetická operace	0
podavač papíru	0
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	0
aritmetická operace	0
podavač papíru	0
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	0
aritmetická operace	0
podavač papíru	0
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 % 50 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	0
podavač papíru	0
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 % 50 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	0
podavač papíru	0
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	0
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	0
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %
83 %	62 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %
83 %	62 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	0
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	1
kulatý stůl	0
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	1
kulatý stůl	1
nová vlna	0
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	1
kulatý stůl	1
nová vlna	1
čerpací stanice	0
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %
70 %	87 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	1
kulatý stůl	1
nová vlna	1
čerpací stanice	1
systém typu	0
centrum města	0
na další	0
program v	0
úroveň být	0

100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %
70 %	87 %
72 %	100 %

Evaluation Metrics

$$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ colloc.|}{|total\ collocations|}$$

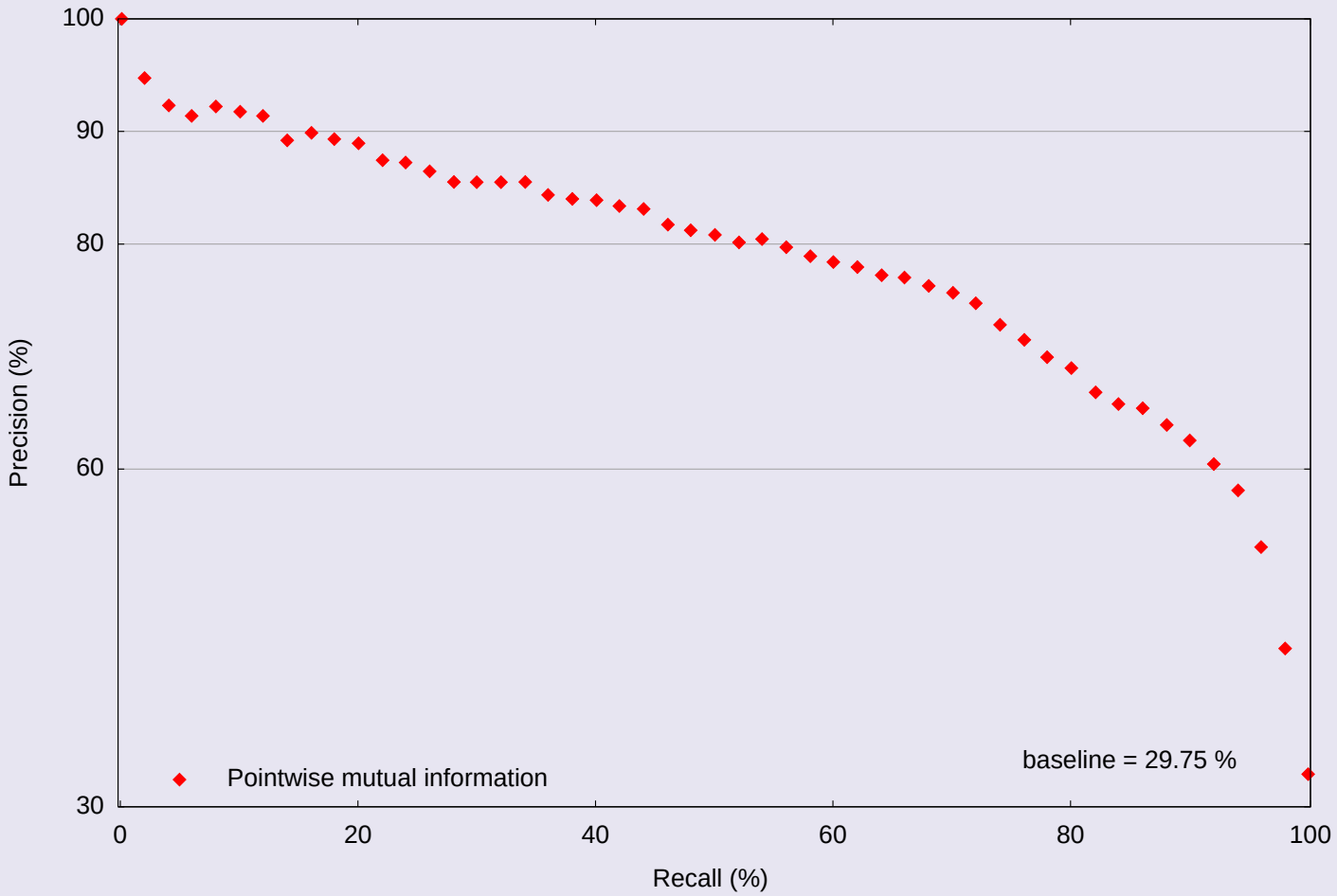
červený kříž	15.66
železná opona	15.23
řádová čárka	14.01
kupónová knížka	13.83
autor knihy	11.05
aritmetická operace	10.52
podavač papíru	10.17
nová kniha	10.09
kulatý stůl	7.03
nová vlna	6.59
čerpací stanice	6.04
systém typu	3.54
centrum města	1.54
na další	0.54
program v	0.35
úroveň být	0.25

červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	1
kulatý stůl	1
nová vlna	1
čerpací stanice	1
systém typu	1
centrum města	1
na další	1
program v	1
úroveň být	1

100 %	12 %
100 %	25 %
100 %	37 %
100 %	50 %
80 %	50 %
83 %	62 %
85 %	75 %
75 %	75 %
77 %	87 %
70 %	87 %
72 %	100 %
66 %	100 %
61 %	100 %
57 %	100 %
53 %	100 %
50 %	100 %

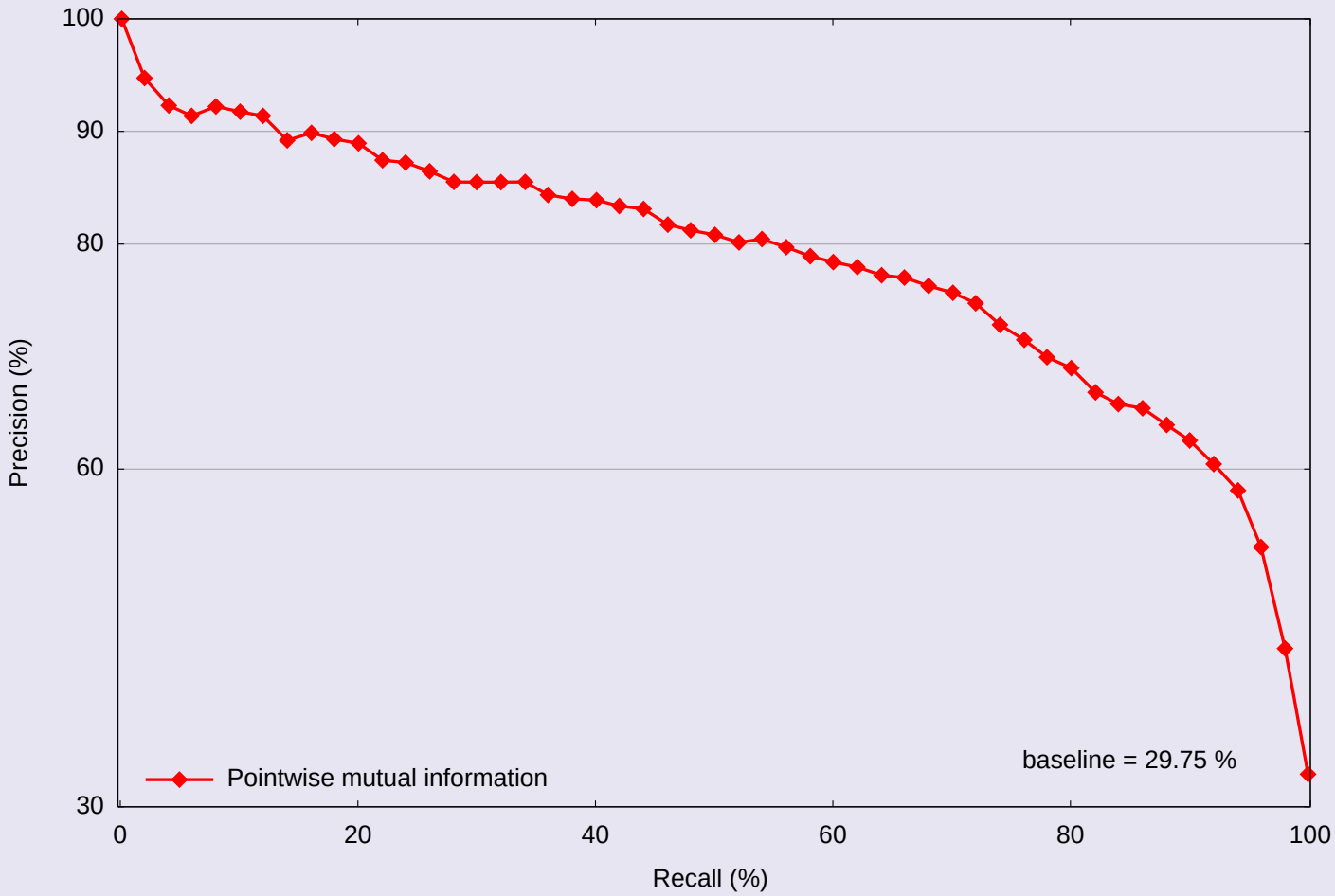
Precision-Recall Graphs

Precision-Recall



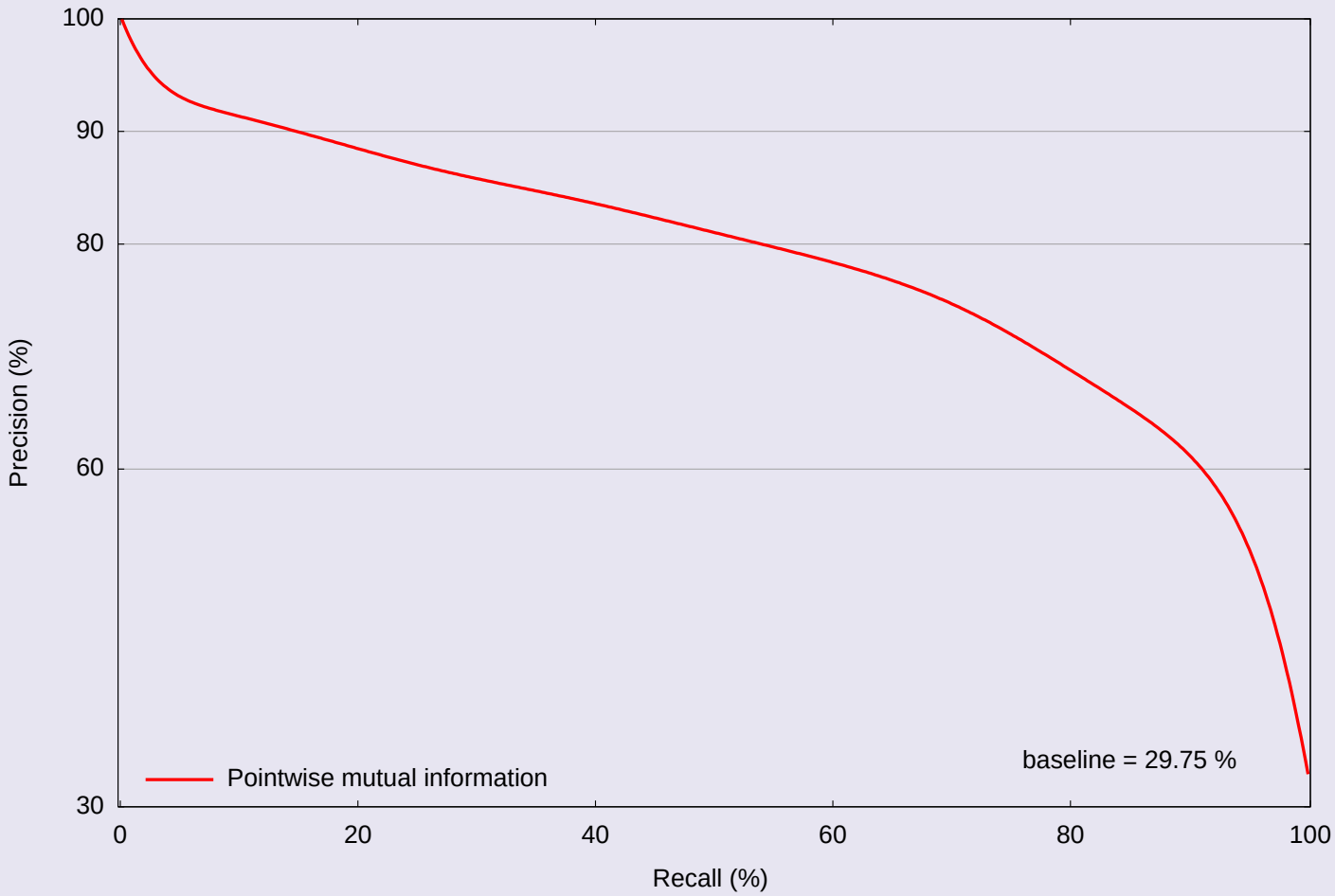
Precision-Recall Graphs

Precision-Recall



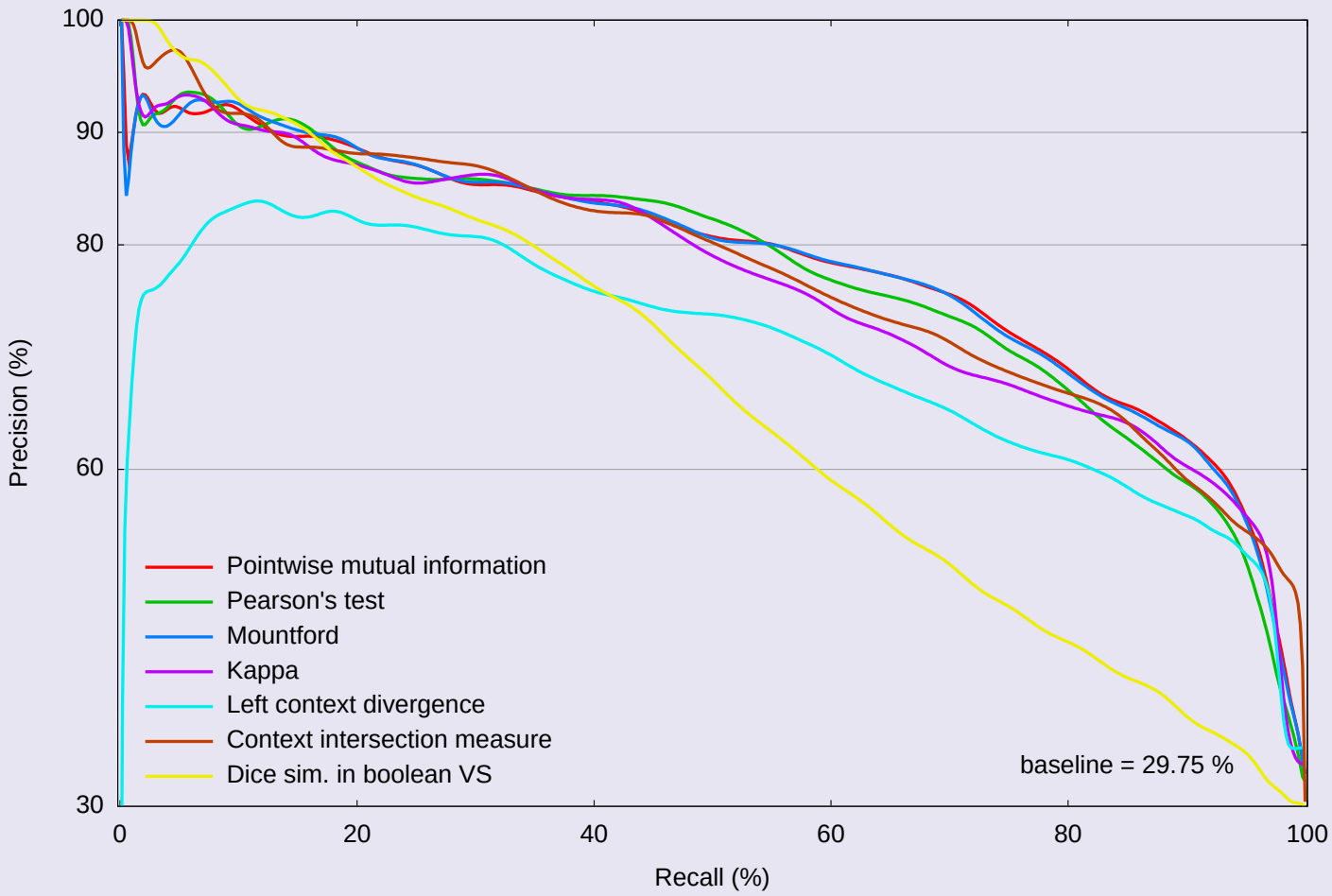
Precision-Recall Graphs

Precision-Recall



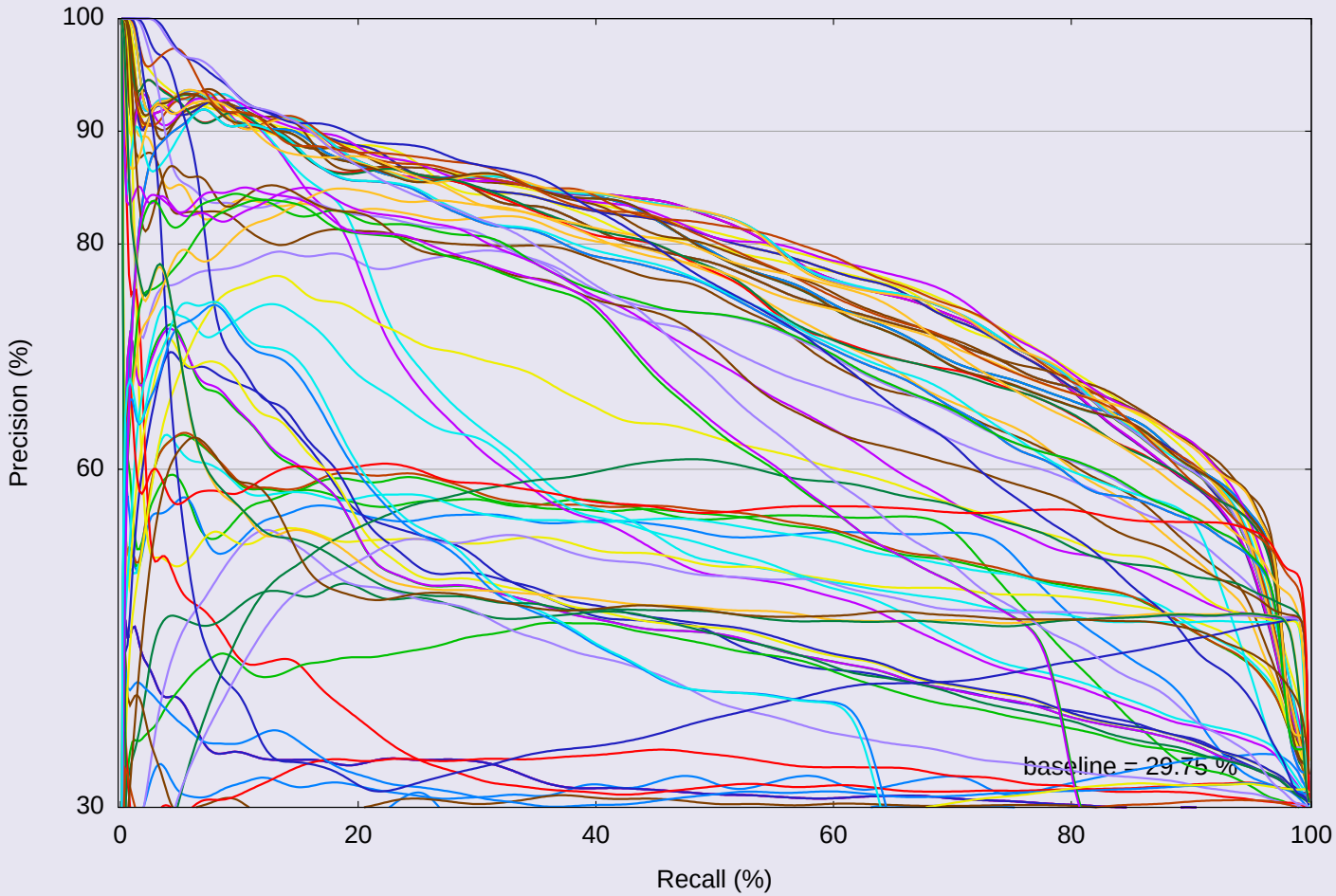
The Best Methods

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



All Results

Precision-Recall curves of all association measures

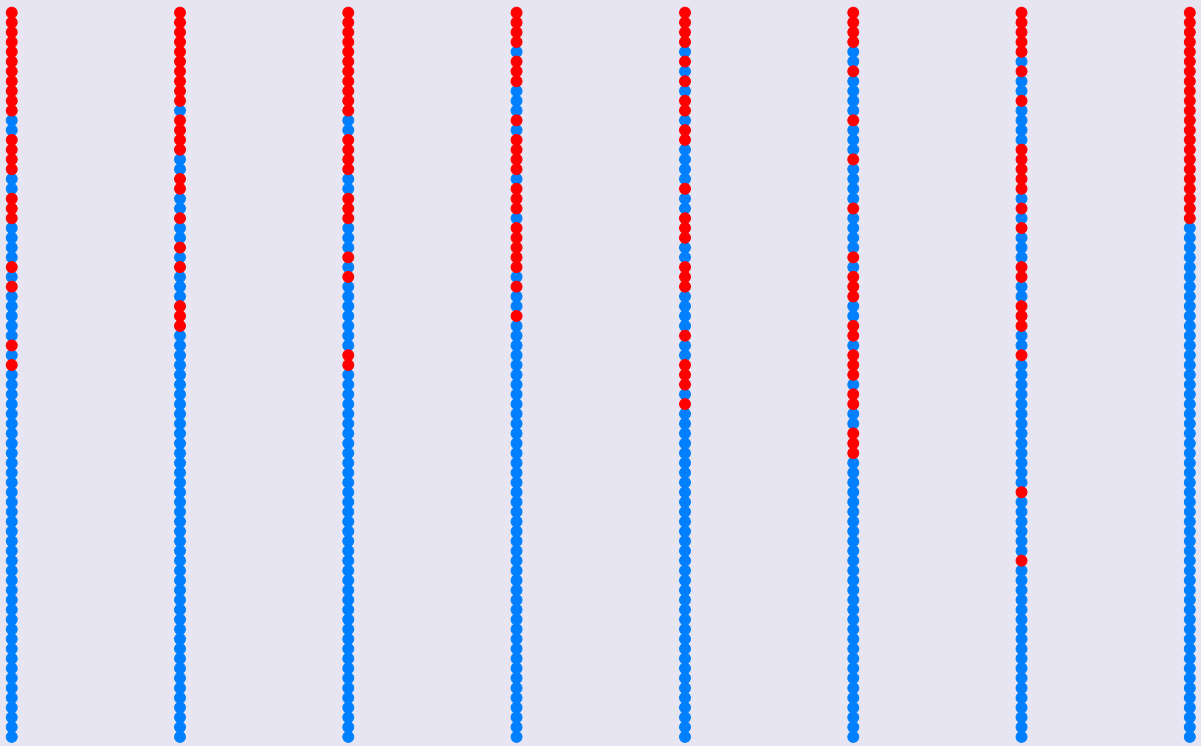


Outline

Motivation I

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

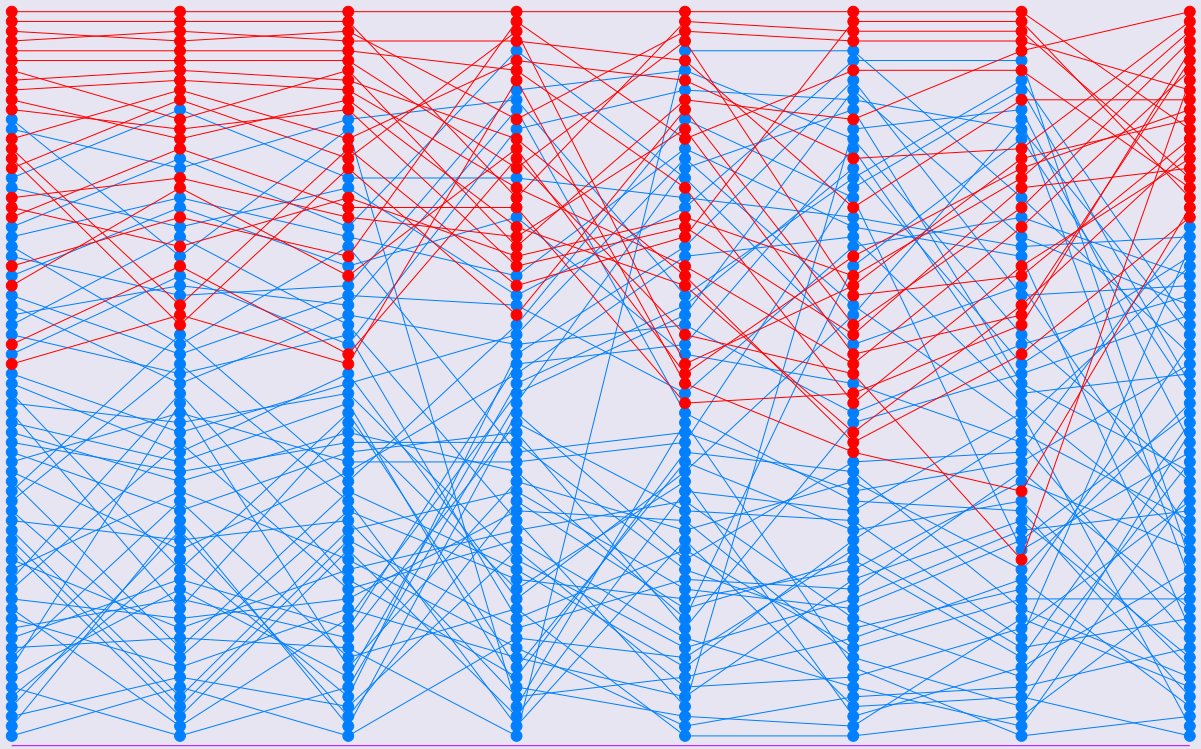
Candidates ranking by different association measures



Motivation I

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

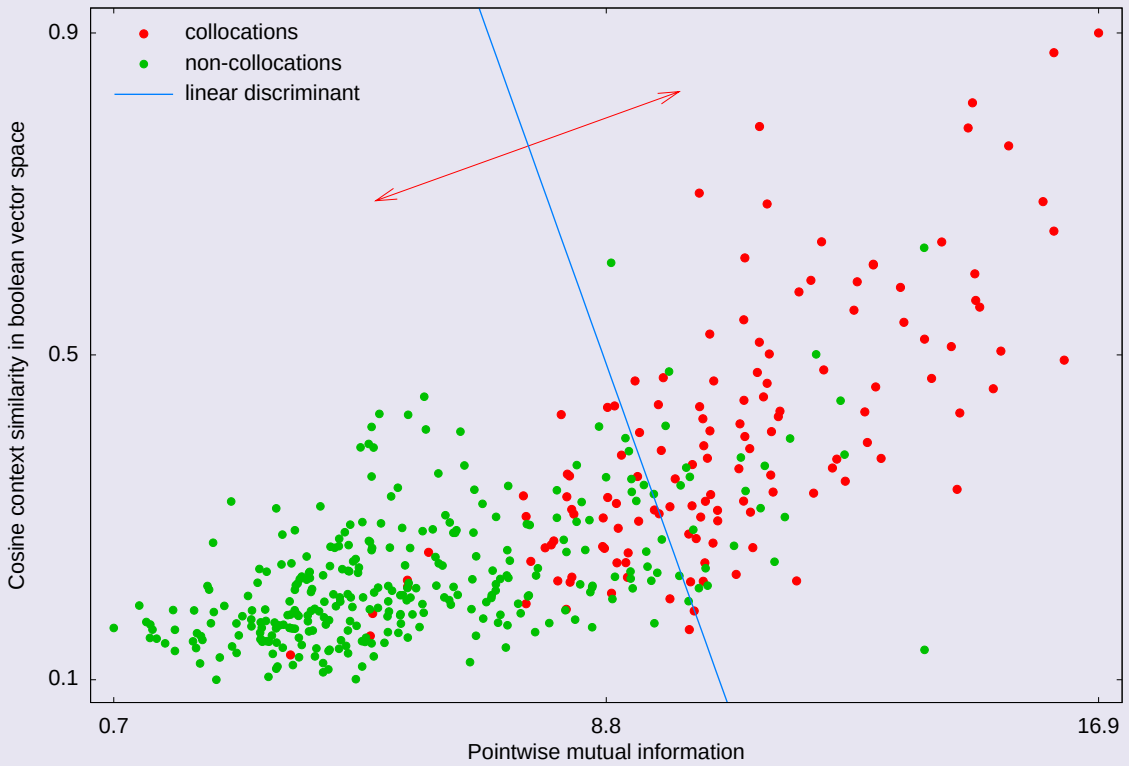
Candidates ranking by different association measures



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 is dependent on the majority of the these votes.

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

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

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

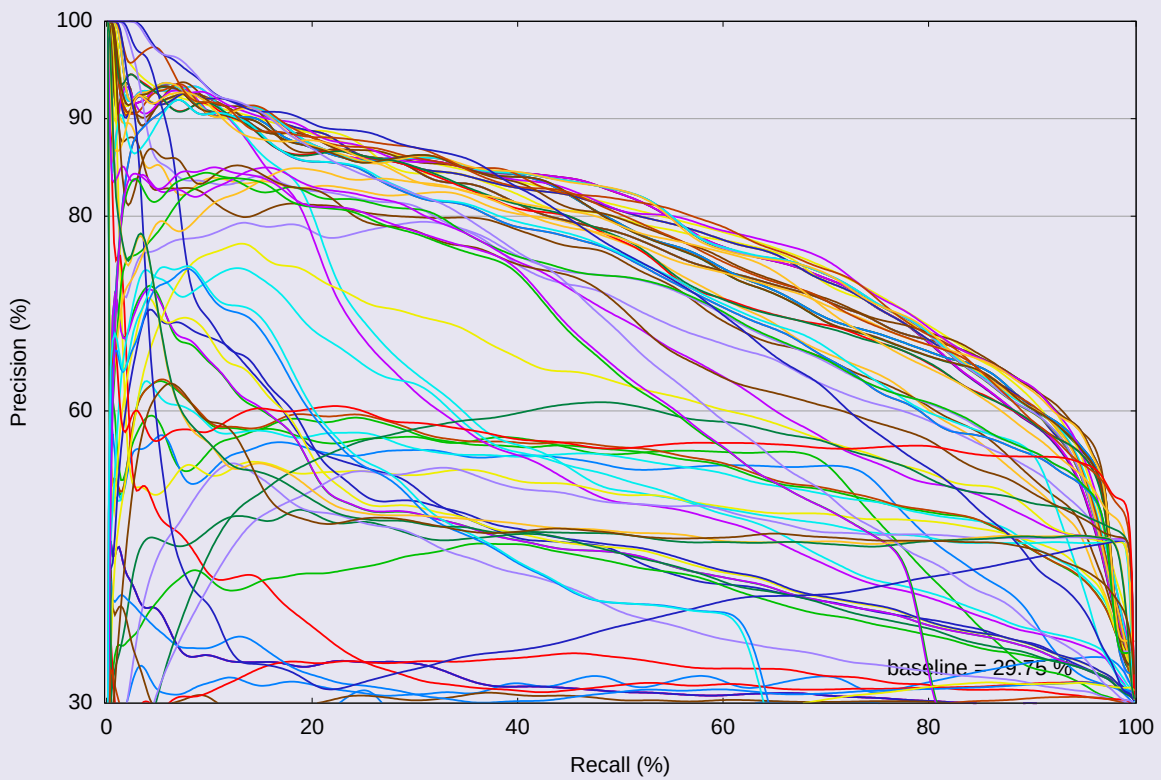
Logistic Regression

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

Logistic Regression

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

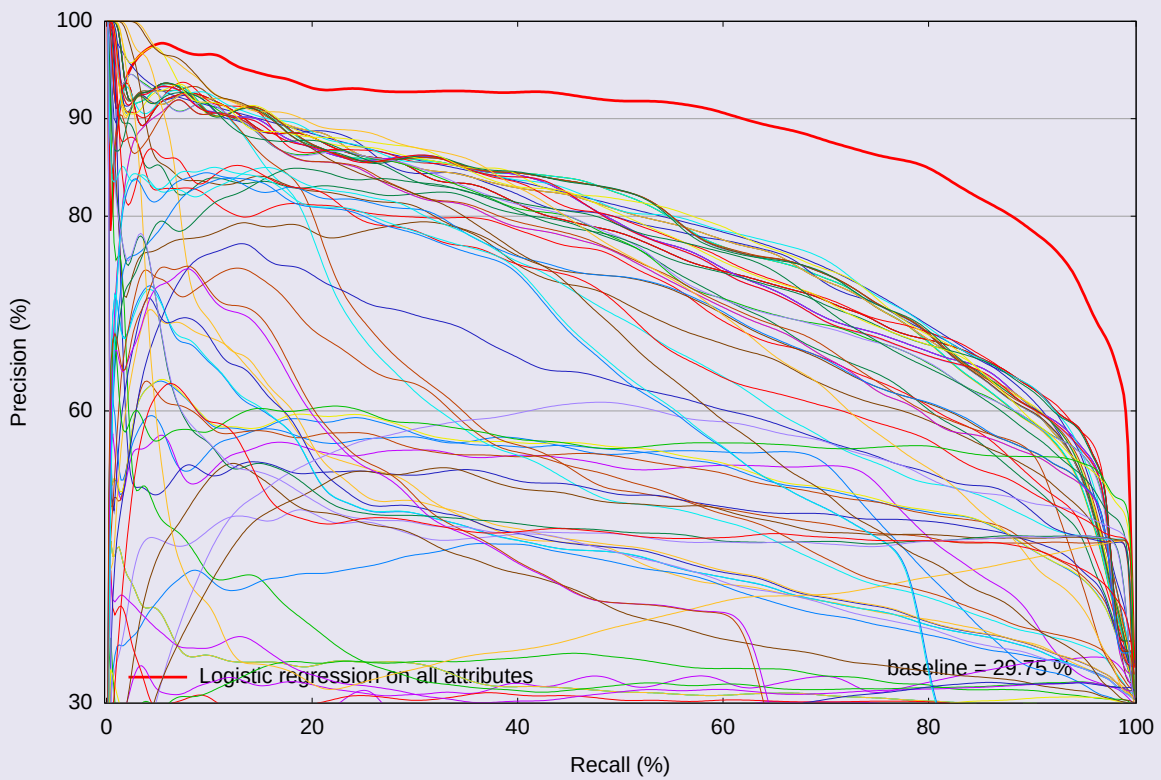
Combination of multiple methods by logistic regression



Logistic Regression

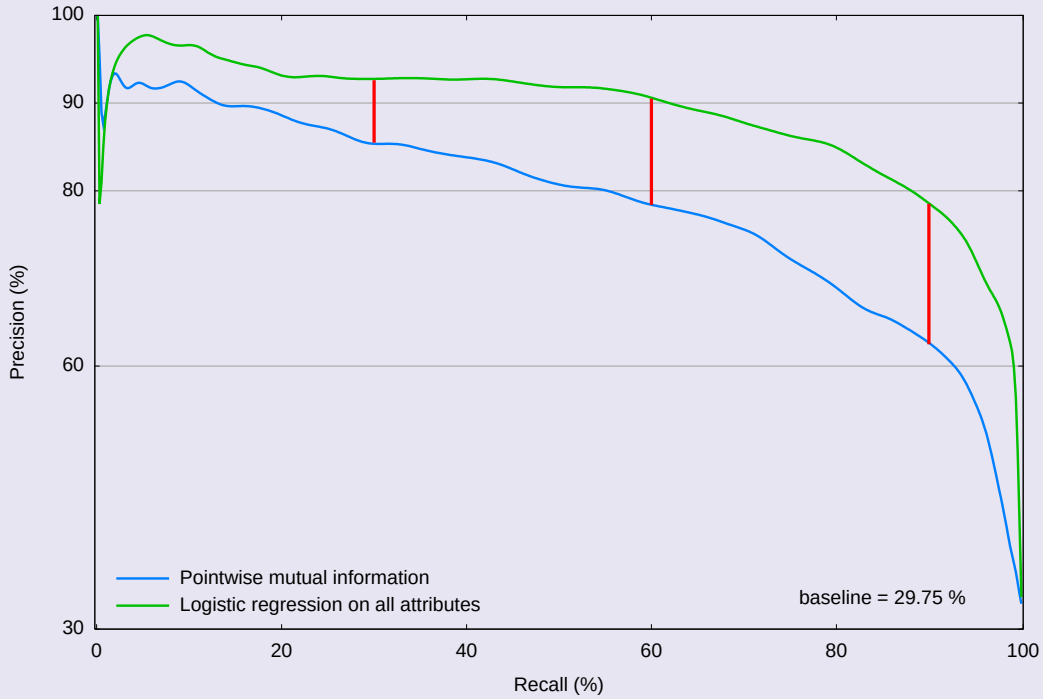
$$P(\mathbf{x}\ is\ collocation) = \frac{\exp^{\beta_0+\beta_1 x_1+\beta_2 x_2 \ldots+\beta_n x_n}}{1+\exp^{\beta_0+\beta_1 x_1 \ldots+\beta_2 x_2+\beta_n x_n}}$$

Combination of multiple methods by logistic regression



Logistic Regression: Results I

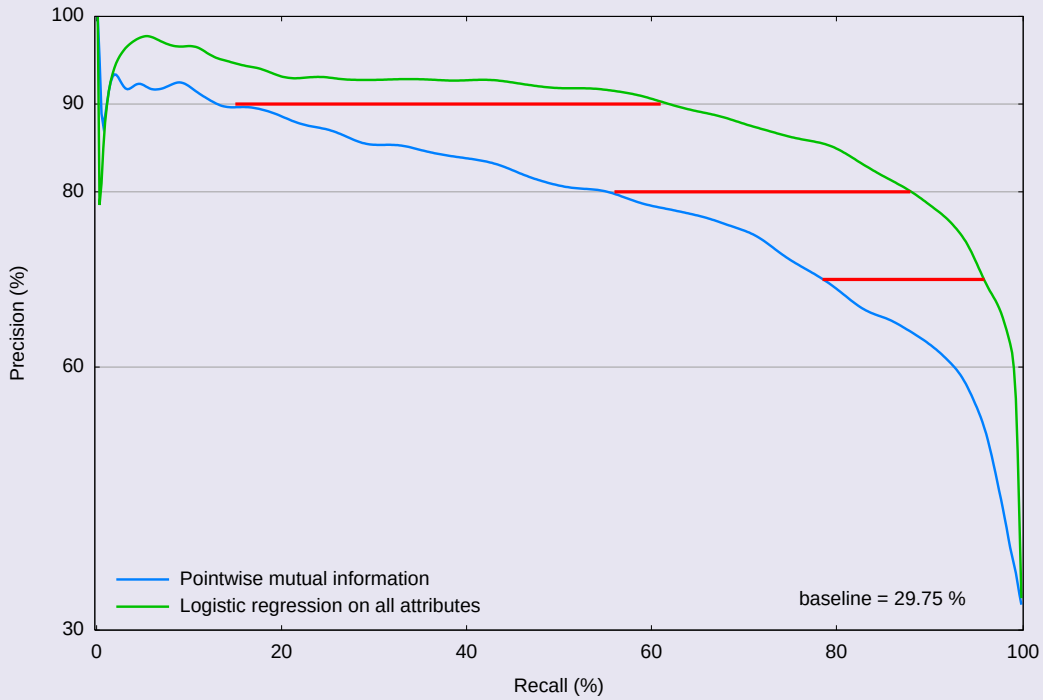
Precision improvement



Recall	30	60	90
P. mutual information	85.5	78.4	62.5
Logistic regression-17	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-17	55.8	86.7	96.7
Absolute improvement	**39.2**	30.7	17.7
Relative improvement	**242.3**	54.8	23.9

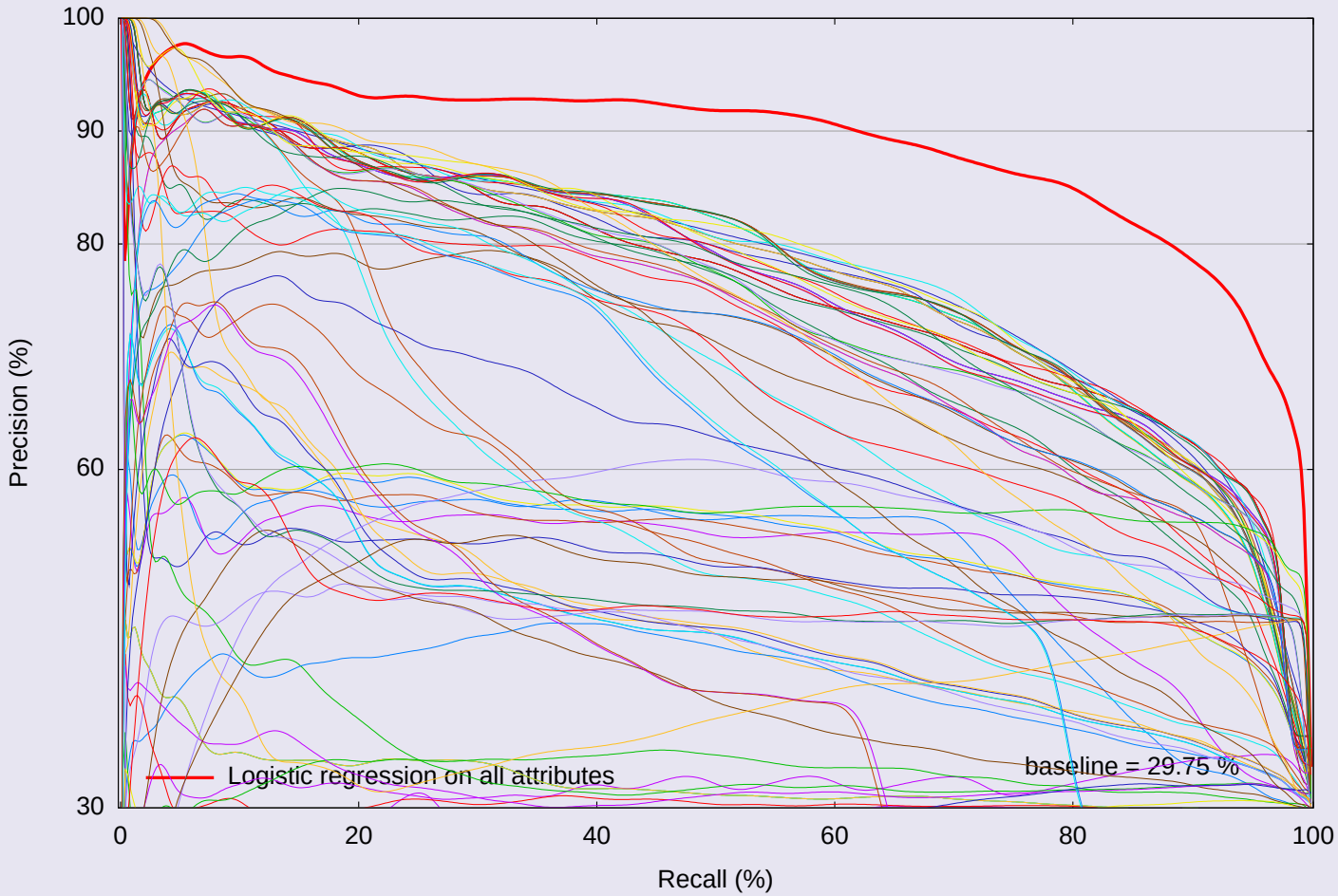
Outline

Attribute Selection

- Can we reduce the number of combined association measures?
- Greedy attribute addition:
  1. Start with an empty set of attributes.
  2. Add the attribute that leads to the best performance improvement.
  3. Repeat until any performance improvement can be achieved.
- Greedy attribute removal:
  1. Start with a full set of attributes.
  2. Remove the attribute that doesn't reduce the performance.
  3. Repeat until any attribute can be removed.

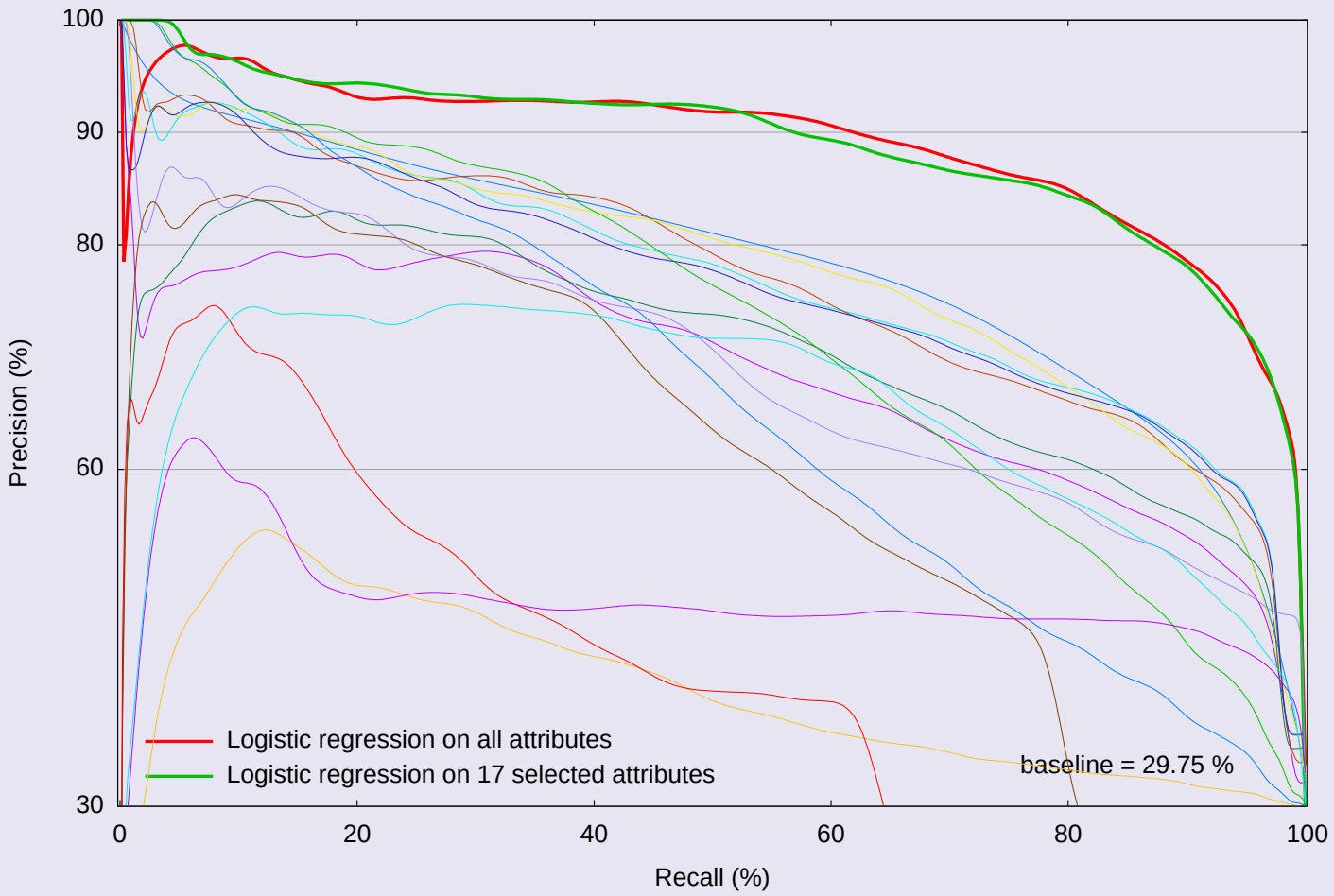
Attribute Selection: Beginning

Logistic regression on all attributes (88 attributes)



Attribute Selection: End

Greedy attribute selection using logistic regression (17 attributes)



Summary

- Achieved results
  - Empirical evaluation of 84 association measures.
    Pointwise mutual information evaluated as the best measure.
  - Statistical combination of multiple association measures.
    Logistic regression leads to significant performance improvement.
  - Selection of the best subset of association measures.
    Greedy algorithm reduced number of association measures to 17.
- Outlook
  - Evaluation experiments on English PennTreebank
  - Perform experiments on a **BIG** data
  - Contribution of collocations in real-world applications

For Further Reading


František Čermák, Jan Holub
Syntagmatika a paradigmatika českého slova: Valence a kolokabilita
SPN, Praha, 1982.


Christipher Manning, Hinrich Schütze
Foundations of Statistical Natural Language Processing
MIT Press, USA, 1999.


Sabine Bartsch
Structural and Functional Properties of Collocations in English
Gunter Narr Verlag Tübingen, Germany, 2004.


Stephan Evert
The statistics of of word cooccurrences: Word Pairs and Collocations
University of Stuttgart, Germany, 2004.


Pavel Pecina
An Extensive Empirical Study on Collocation Extraction Methods
Proceedings of the ACL'05 Student Research Workshop, USA, 2005.

That's all folks . . .

Thank you!