Lexical Association Measures

Collocation Extraction

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Charles University, Prague

Doctoral thesis defense
Prague, September 24, 2008

Presentation outline

1. Introduction
2. Collocation extraction
3. Lexical association measures
4. Reference data
5. Empirical evaluation
6. Combining association measures
7. Conclusions

Lexical association

Collocational association

- restricts combination of words into phrases (beyond grammar)
- *collocations*

  crystal clear, cosmetic surgery, weapons of mass destruction

Semantic association

- reflects semantic relationship between words
- *synonymy, antonymy, hypo/hyperonymy, meronymy, etc.*

  sick – ill, baby – infant, dog – cat

Cross-language association

- corresponds to potential translations of words between languages
- *translation equivalents*

  maison – house, baum – tree, květina – flower

Lexical association

Collocational association

- restricts combination of words into phrases (beyond grammar)
- *collocations* → *lexicon*

  crystal clear, cosmetic surgery, weapons of mass destruction

Semantic association

- reflects semantic relationship between words
- *synonymy, antonymy, hypo/hyperonymy, meronymy, etc.* → *thesaurus*

  sick – ill, baby – infant, dog – cat

Cross-language association

- corresponds to potential translations of words between languages
- *translation equivalents* → *dictionary*

  maison – house, baum – tree, květina – flower

Measuring lexical association

Motivation

- automatic acquisition of associated words (*into a lexicon/thesarus/dictionary*)

Tool: Lexical association measures

- mathematical formulas determining strength of association between two (or more) words based on their occurrences and cooccurrences in a corpus.

Applications

- lexicography, natural language generation, word sense disambiguation
- bilingual word alignment, identification of translation equivalents
- information retrieval, cross-lingual information retrieval
- keyword extraction, named entity recognition
- syntactic constituent boundary detection, dependency parsing
- collocation extraction

Goals, objectives, and limitations

Goal

- application of lexical association measures to collocation extraction

Objectives

1. to compile a comprehensive inventory of lexical association measures
2. to build several reference data sets for collocation extraction
3. to evaluate the lexical association measures on these data sets
4. to explore the possibility of combining these measures into more complex models and improve performance in collocation extraction

Limitations

- ✘ focus on bigram (*two-word*) collocations

Collocational association

Collocability

- the ability of words to combine with other words in text
- governed by a system of rules and constraints: *syntactic, semantic, pragmatic*
- must be adhered to in order to produce correct, meaningful, fluent utterances
- specified *intensionally* or *extensionally*

Collocations

- extensionally restricted word combinations
- should be listed in a lexicon and learned in the same way as single words

Types of collocations

1. idioms *(to kick the bucket, to hear st. through the grapevine)*
2. proper names *(New York, Old Town)*
3. technical terms *(car oil, stock owl)*
4. phrasal verbs *(to switch off, to look after)*
5. light verb compounds *(to take a nap, to do homework)*
6. lexically restricted expressions *(strong tea, broad daylight)*

Collocation extraction

Task

- to extract a list of collocations (*types*) from a text corpus (collocation lexicon)
- no need to identify particular occurrences (*instances*) of collocations

Methods

- based on extraction principles verifying characteristic collocation properties
- i.e. hypotheses about word occurences and cooccurrences in the corpus
- formulated as lexical association measures
- compute association score for each collocation candidate from the corpus
- the scores indicate a chance of a candidate to be a collocation

Extraction principles

1. *"Collocation components occur together more often than by chance"*
2. *"Collocations occur as units in information-theoretically noisy environment"*
3. *"Collocations occur in different contexts to their components"*

Extraction pipeline

1. linguistic preprocessing (*morphological and syntactic level*)
2. identification of collocation candidates (*dependency/surface/distance bigrams*)
3. extraction of occurrence and cooccurrence statistics (*frequency, contexts*)
4. filtering the candidates to improve precision (*POS patterns, frequency*)
5. application of a choosen lexical association measure
6. ranking/classification of collocation candidates according to their scores

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*

Lexical association measures

#	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**	$-2\sum_{i,j} f_{ij} \log f_{ij}/\hat{f}_{ij}$
16.	**Squared log likelihood ratio**	$-2\sum_{i,j} \log f_{ij}^2/\hat{f}_{ij}$
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**	$\frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$
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**	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$

#	Name	Formula
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**	$\sqrt{P(xy)} \cdot AV$
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**	$P(*y) \log P(*y) - \sum_w P(w\|C^r_{xy}) \log P(w\|C^r_{xy})$
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**	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sqrt{\sum_w P(w\|C_x)^2 \cdot \sum_w P(w\|C_y)^2}}$
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 pointwise *MI***	$\frac{\sum_w MI(w,x)MI(w,y)}{\sqrt{\sum_w MI(w,x)^2} \cdot \sqrt{\sum_w MI(w,y)^2}}$
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**	$z_i = f(w_i\|C_z) \cdot \frac{N}{df(w_i)};\ df(w_i) = \|\{x : w_i \epsilon C_x\}\|$
	Dice context similarity:	$\frac{1}{2}(\text{dice}(\mathbf{c}_x, \mathbf{c}_{xy}) + \text{dice}(\mathbf{c}_y, \mathbf{c}_{xy}))$ $\mathbf{c}_z = (z_i);\ \text{dice}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2\sum x_i y_i}{\sum {x_i}^2 + \sum {y_i}^2}$
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**	$z_i = f(w_i\|C_z) \cdot \frac{N}{df(w_i)};\ df(w_i) = \|\{x : w_i \epsilon C_x\}\|$

Table 1: Inventory of lexical association measures for collocation extraction.

Reference data set: *PDT-Dep*

Source corpus

- Prague Dependency Treebank 2.0, 1.5 mil. tokens
- manually annotated on *morphological* and *analytical* level

Collocation candidates

- dependency bigrams: direct dependency relation between components
- morphological normalization: *lemma proper + pos + gender + degree + negation*
- part-of-speech filter: *A:N, N:N, V:N, R:N, C:N, N:V, N:C, D:A, N:A, D:V, N:T, N:D, D:D*
- frequency filter: *minimal frequency required,* $f > 5$

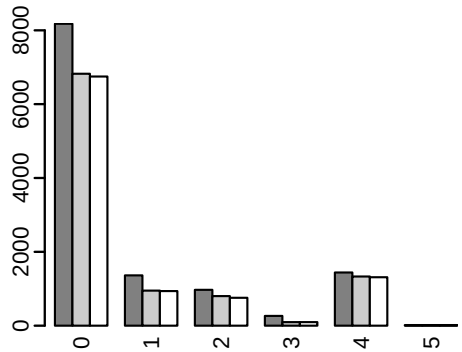
Annotation

- three independent parallel annotations (*no context; full agreement required*)
- six categories, merged into two: collocations (1-5), non-collocations (0):

5. *idiomatic expressions*
4. *technical terms*
3. *support verb constructions*
2. *proper names*
1. *stock phrases, frequent unpredictable usages*

---

0. *non-collocations*



Additional data sets

PDT-Surf

- analogous to *PDT-Dep* (*corpus, filtering, annotation*)
- collocation candidates extracted as surface bigrams: pairs of adjacent words
- assumption: collocations cannot be modified by insertion of another word
- annotation consistent with *PDT-Dep*

CNC-Surf

- collocation candidates – instances of *PDT-Surf* in the *Czech National Corpus*
- SYN 2000 and 2005, 240 mil. tokens, morphologicaly tagged and lemmatized

PAR-Dist

- source corpus: Swedish Parole, 22 mil. tokens
- automatic morphological tagging and lemmatization
- distance bigrams: word pairs occurring within a distance of 1–3 words
- annotation: non-exhaustive manual extraction of support verb constructions
- no frequency filter applied

Reference data summary

reference data set	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*	*PAR-Dist*
morphology	*manual*	*manual*	*auto*	*auto*
syntax	*manual*	*none*	*none*	*none*
bigram types	*dependency*	*surface*	*surface*	*distance*
tokens	1 504 847	1 504 847	242 272 798	22 883 361
bigram types	635 952	638 030	30 608 916	13 370 375
after frequency filtering	26 450	29 035	2 941 414	13 370 375
after part-of-speech filtering	12 232	10 021	1 503 072	898 324
collocation candidates	12 232	10 021	9 868	17 027
data sample size	100 %	100 %	0.66 %	1.90 %
true collocations	2 557	2 293	2 263	1 292
baseline precision (%)	21.02	22.88	22.66	7.59

Evaluation measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Evaluation measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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 measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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 measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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

Classification	
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 measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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

Classification	
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 measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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

Classification	
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

Precision	Recall
100 %	50 %

Evaluation measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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

Classification	
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

Precision	Recall
100 %	50 %
80 %	50 %

Evaluation measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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

Classification	
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

Precision	Recall
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 %

Evaluation measures

1) $$Precision = \frac{|correctly\ classified\ collocations|}{|total\ classified\ as\ collocations|} \qquad Recall = \frac{|correctly\ classified\ collocations|}{|total\ collocations|}$$

Ranking	
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

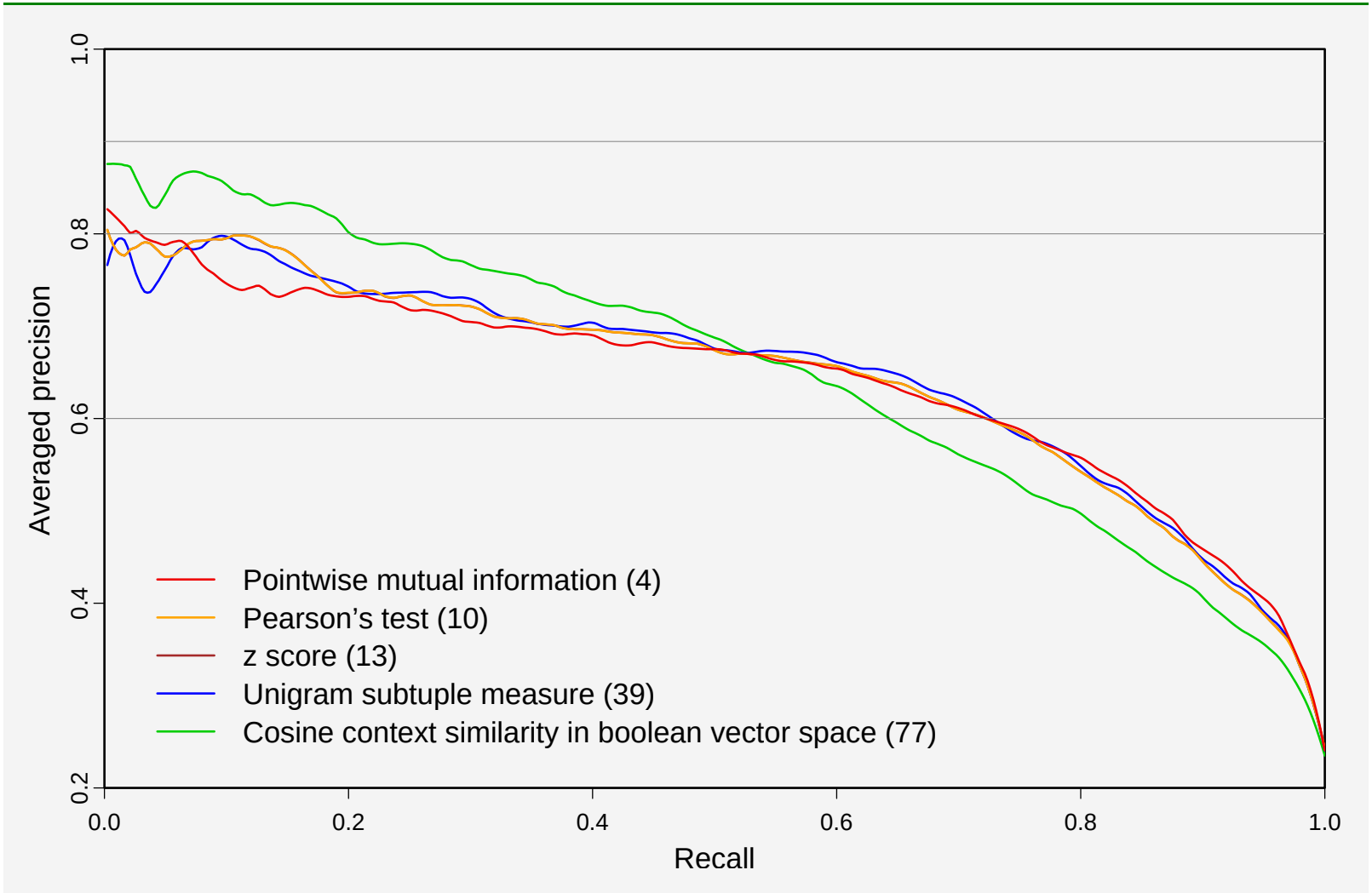
Classification	
červený kříž	1
železná opona	1
řádová čárka	1
kupónová knížka	1
autor knihy	1
aritmetická operace	1
podavač papíru	1
nová kniha	1
kulatý stůl	1
nová vlna	1
čerpací stanice	1
systém typu	1
centrum města	1
na další	1
program v	1
úroveň být	1

Precision	Recall
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 %

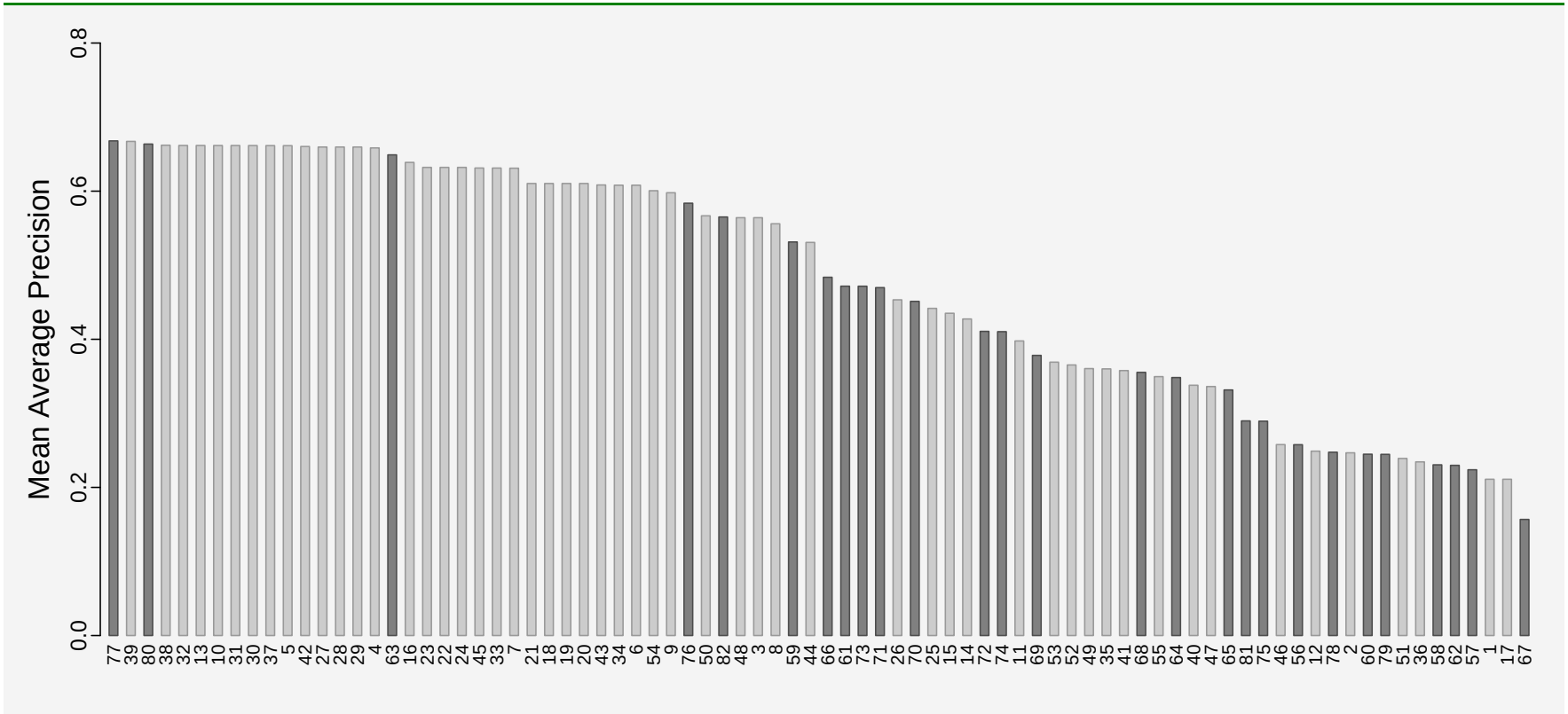
2) $$(Mean)\ Average\ Precision = \frac{1}{r}\sum_{i=1}^{r} p_i$$

89.6 % = AP

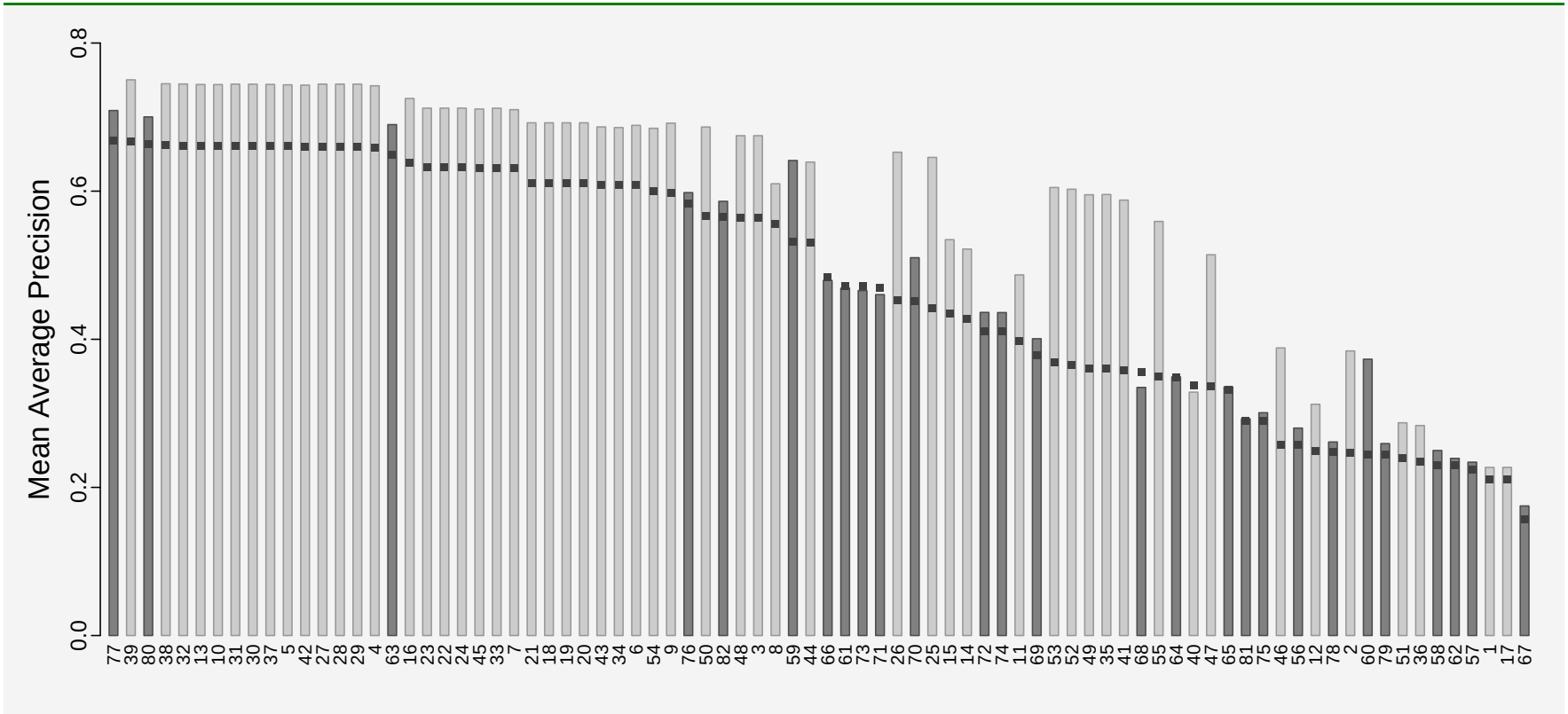
Results / Precision-Recall: *PDT-Dep*



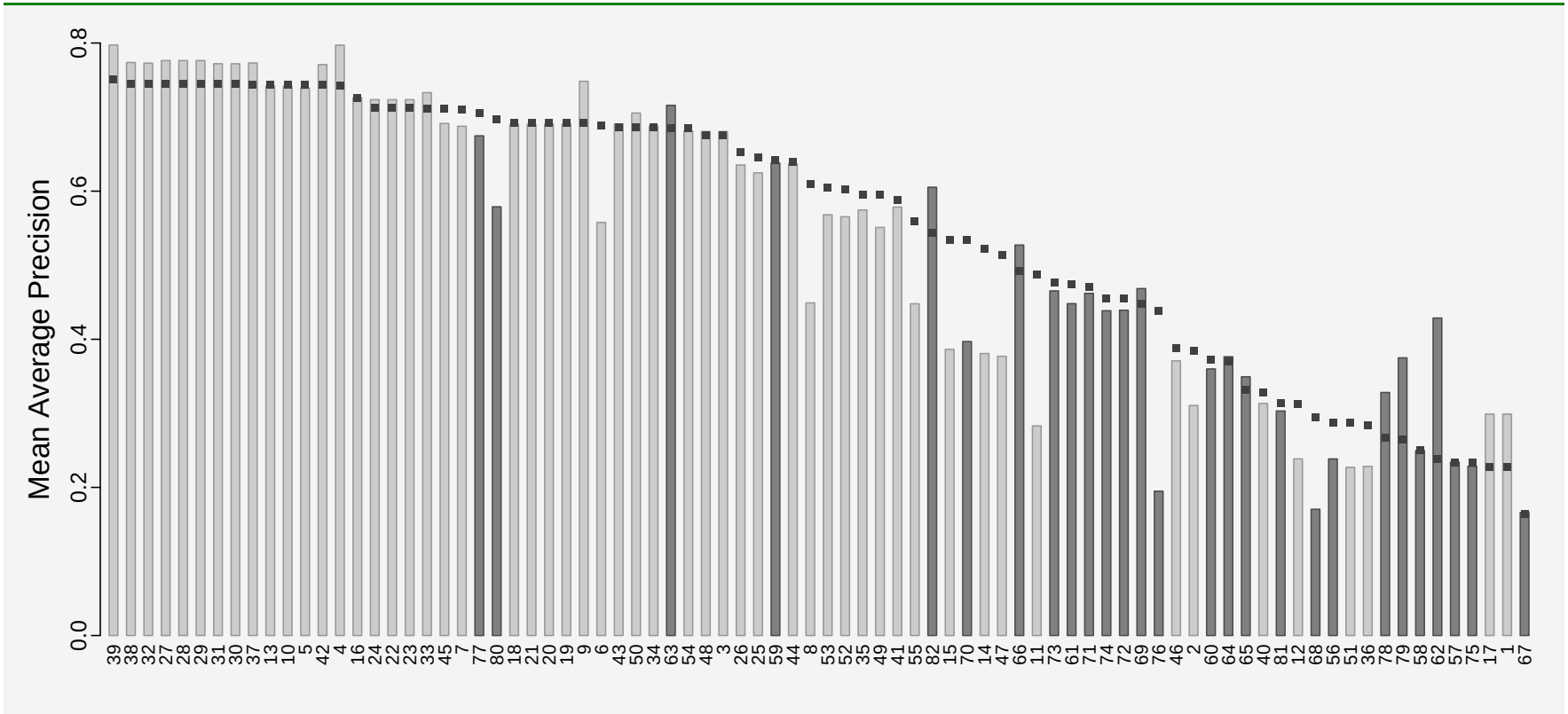
Results / Mean average precision: *PDT-Dep*



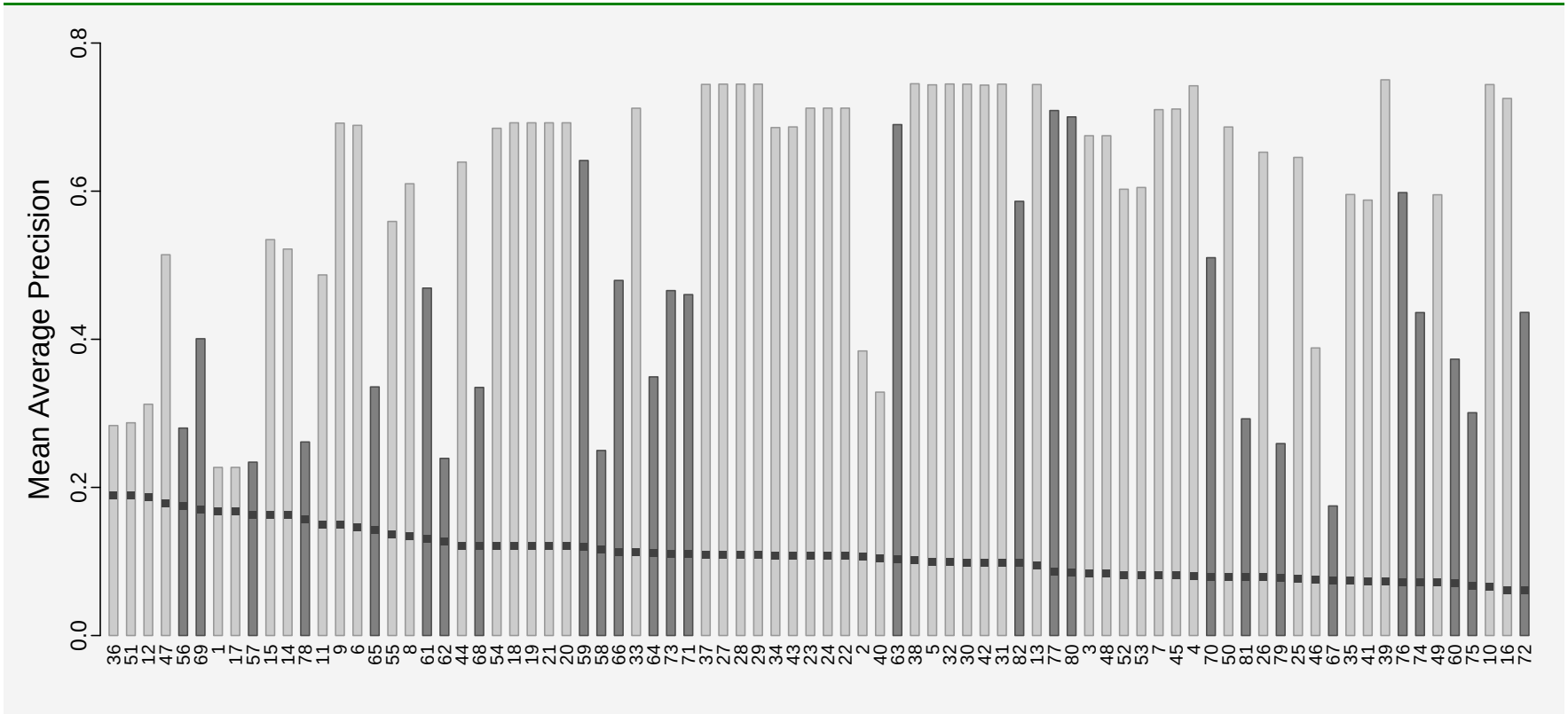
Results / Mean average precision: *PDT-Dep* vs. *PDT-Surf*



Results / Mean average precision: *PDT-Surf* vs. *CNC-Surf*



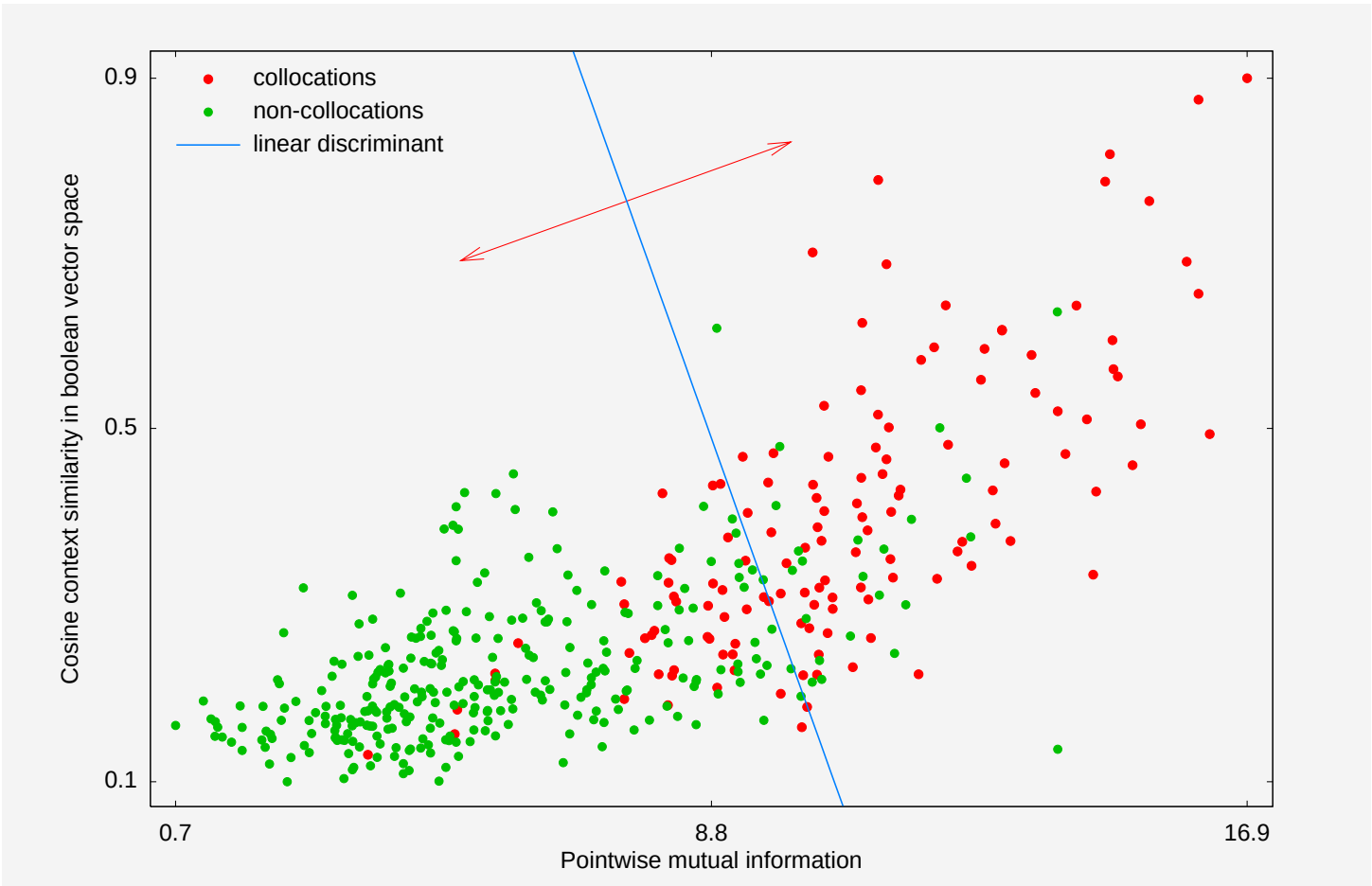
Results / Mean average precision: *PAR-Dist* vs. *PDT-Dep*



Combining association measures

Motivation

- different association measures discover different groups/types of collocations
- existence of uncorrelated association measures



Combination models

Framework

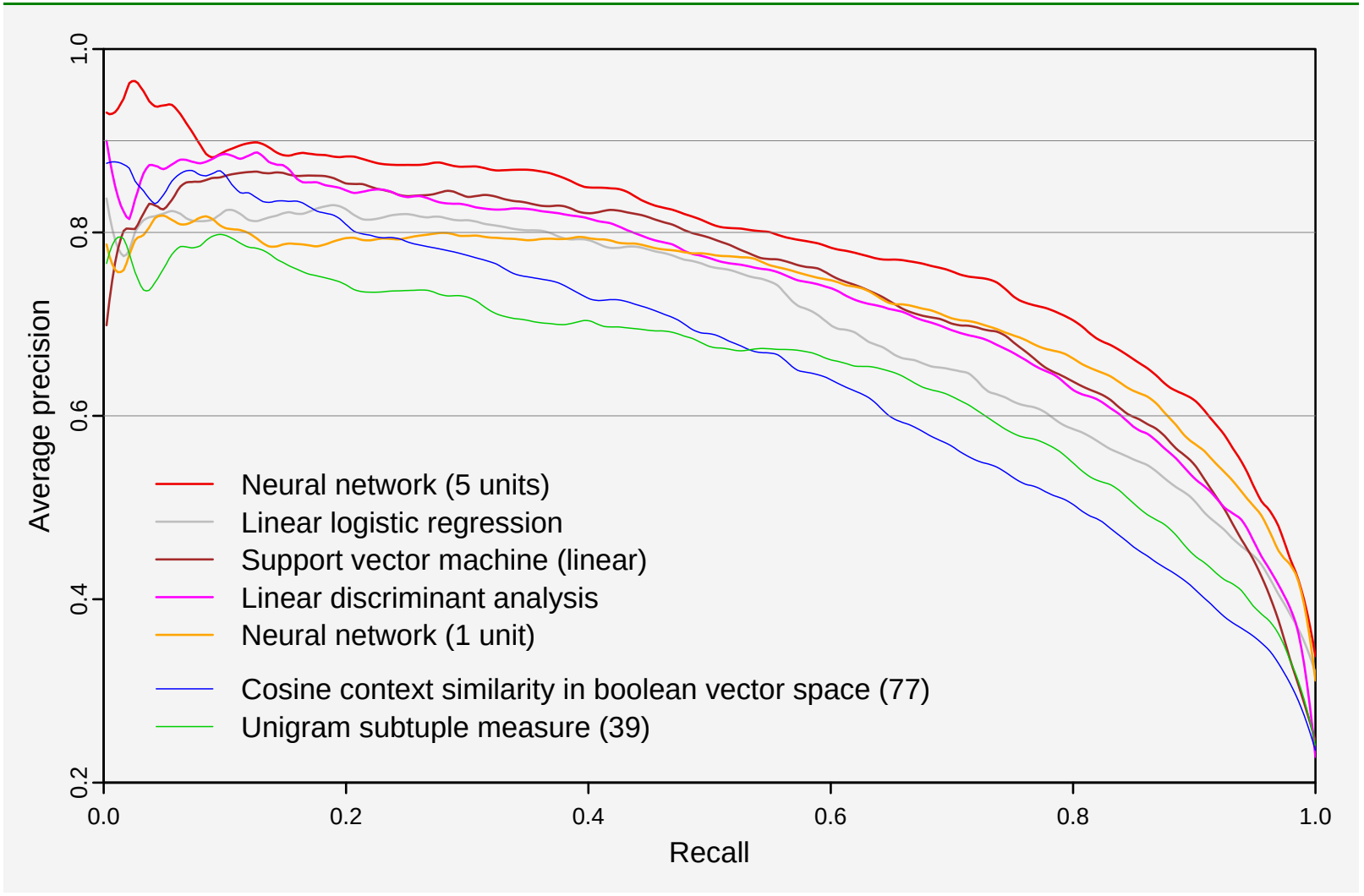
- each collocation candidate x^i is described by the feature vector $\mathbf{x}^i = (x_1^i, \ldots, x_{82}^i)^T$ consisting of scores of all association measures
- and assigned a label $y^i \in \{0, 1\}$ indicating whether the bigram is considered to be a true collocation ($y = 1$) or not ($y = 0$)
- we look for a ranker function $f(\mathbf{x}^i)$ determining the strength of lexical association between components of a candidate x^i
- e.g. linear combination of association scores: $f(\mathbf{x}^i) = w_0 + w_1 x_1^i + \ldots + w_{82} x_{82}^i$

Methods

1. *Linear logistic regression*
2. *Linear discriminant analysis*
3. *Support vector machines*
4. *Neural networks*

- in the training phase used as regular classifiers on two-class data
- in the application phase no classification threshold applies

Results / Precision-Recall: *PDT-Dep*



Results / Mean average precision: *PDT-Dep*

method	*precision at*				
	R=20	*R=50*	*R=80*	*MAP*	*+%*
Neural Network (5 units)	91.00	81.75	70.22	80.87	21.08
Linear Logistic Regression	86.96	79.74	64.63	77.36	15.82
Linear Discriminant Analysis	85.99	77.34	61.44	75.16	12.54
Neural Network (1 unit)	82.47	77.08	65.75	74.88	12.11
Support Vector Machine	81.33	76.08	61.49	73.03	9.35
Cosine similarity (77)	80.88	68.46	49.99	66.79	0.00
Unigram subtuples (39)	75.86	68.19	55.13	66.72	–

Adding linguistic features

Combined variables:

AM: association measures

AM+POS: association measures + part-of-speech pattern

AM+POS+DEP: association measures + part-of-speech + analytical function

reference data set	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*	*PAR-Dist*
Baseline	21.02	22.88	22.66	7.59
Best AM	66.72	75.03	79.74	18.88
NNet.5 (AM)	80.87	84.35	86.30	35.78
NNet.5 (AM+POS)	82.79	86.48	88.22	–
NNet.5 (AM+POS+DEP)	84.53	–	–	–

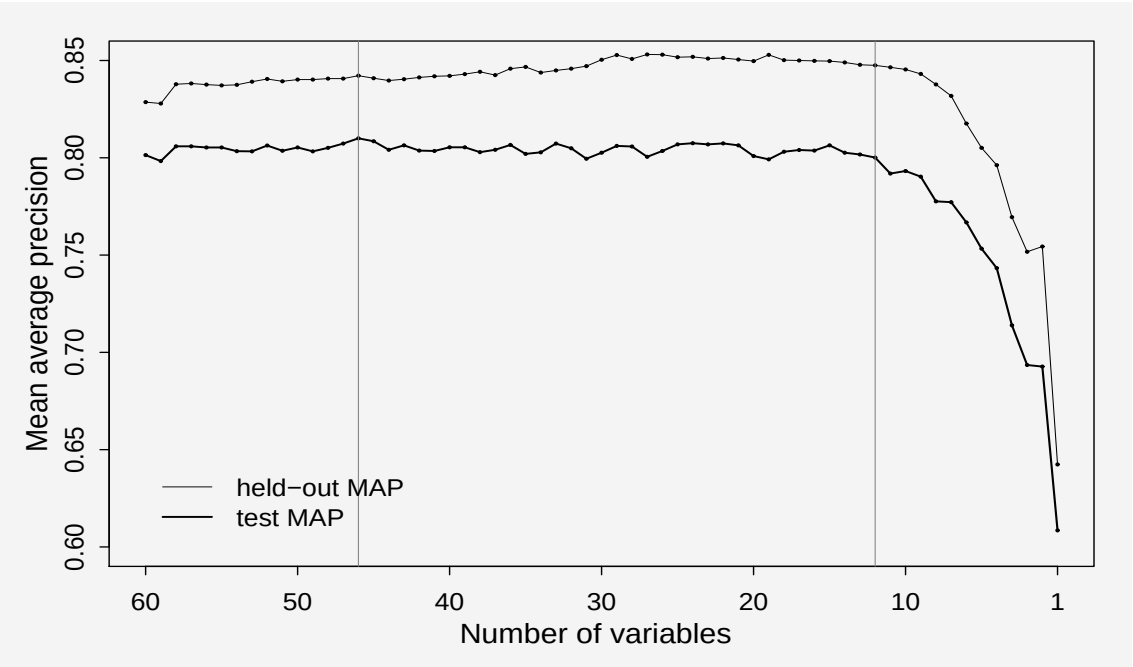
Model reduction

Motivation

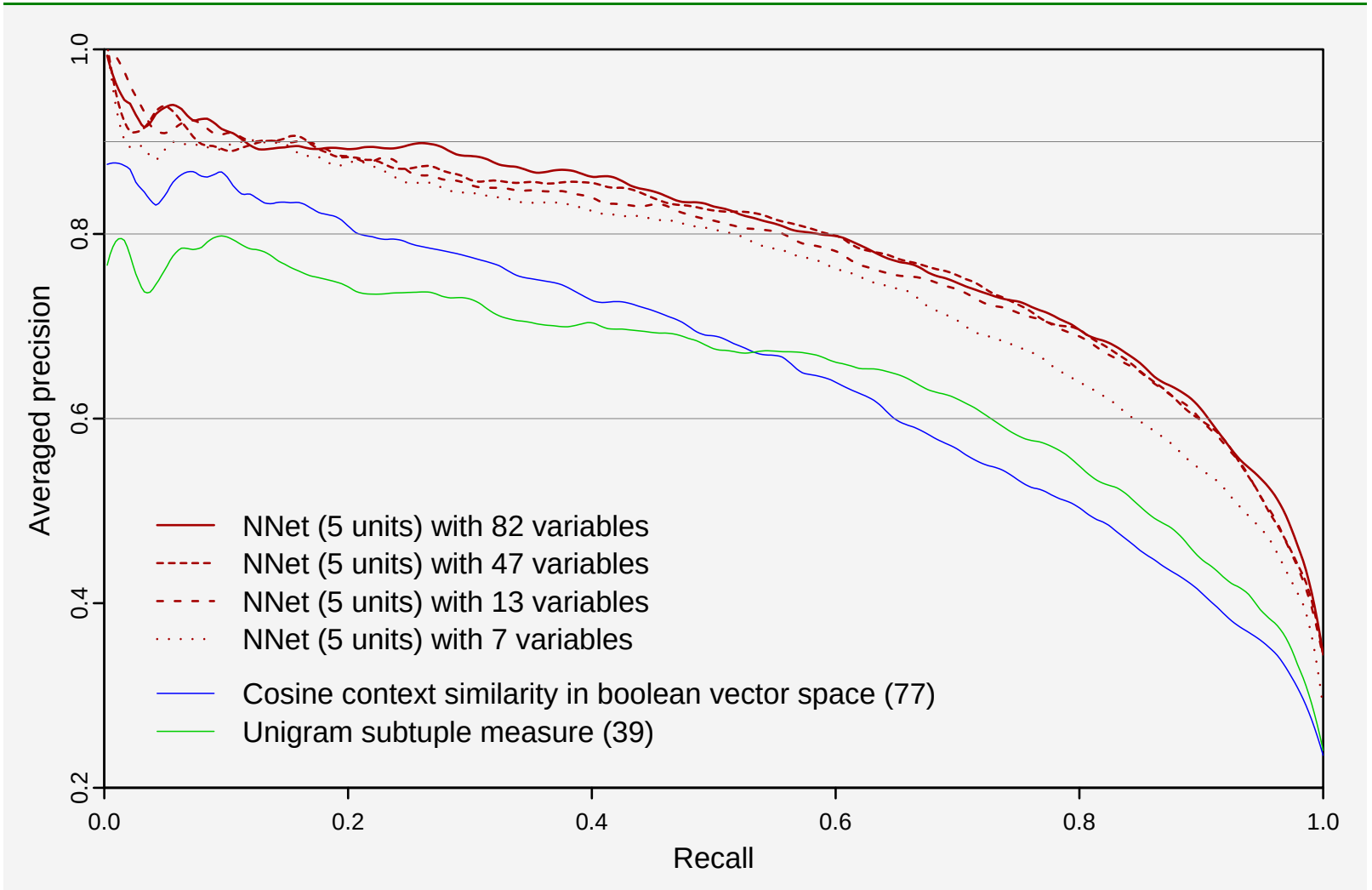
- models combining all 82 association measures are too complex
- redundant/improper variables should be removed

Algorithm

1. correlation based clustering; one representative selected from each cluster
2. step-wise procedure removing the variable causing minimal performance degradation in each iteration



Model reduction results / Precision-Recall: *PDT-Dep*



Conslusions

Objectives accomplished

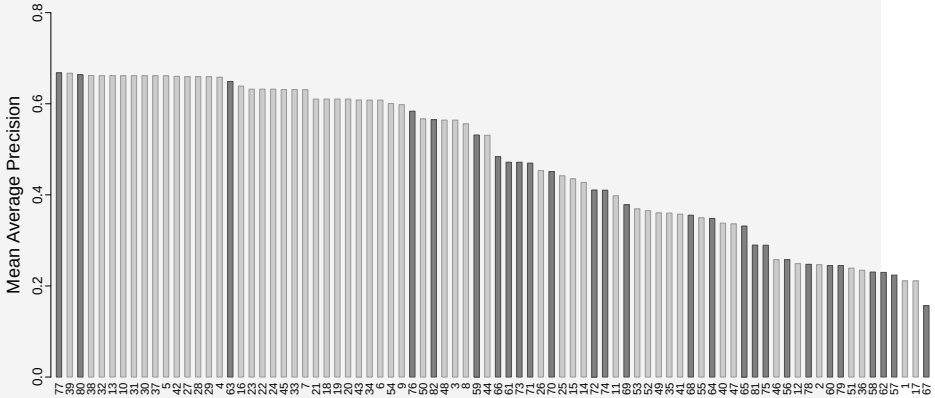
1. to compile a comprehensive inventory of lexical association measures
   - *82 association measures based on 3 extraction principles, unified notation*
2. to build several reference data sets for collocation extraction
   - *4 reference data sets for evaluation in different settings*
3. to evaluate the lexical association measures on these data sets
   - *performance heavily depends on the task, data, language, etc.*
   - *no general "best-performing" measure selected*
4. to combine these measures and improve performance in collocation extraction
   - *combining is meaningfull and greatly improves performance*
   - *combination models can be reduced to use only 13 variables*

Other contribution

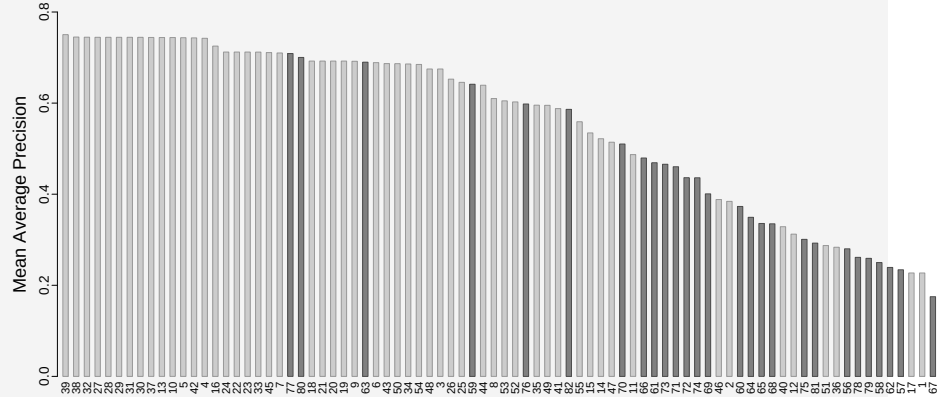
- overview of different notions of collocation (*definitions, typology, classification*)
- evaluation scheme (*average precision, crossvalidation, significance tests*)
- reference data sets used in MWE 2008 Shared Task

Context-based vs. statistical association measures

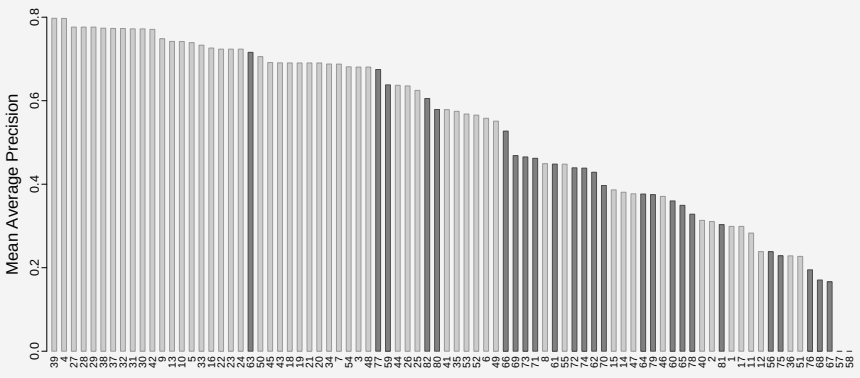
PDT-Dep



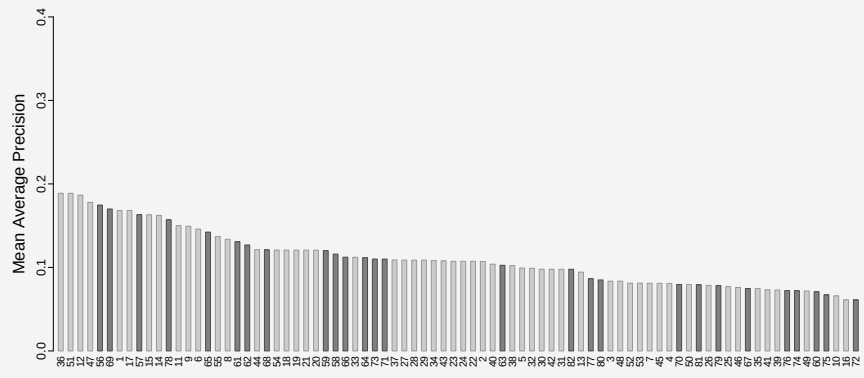
PDT-Surf



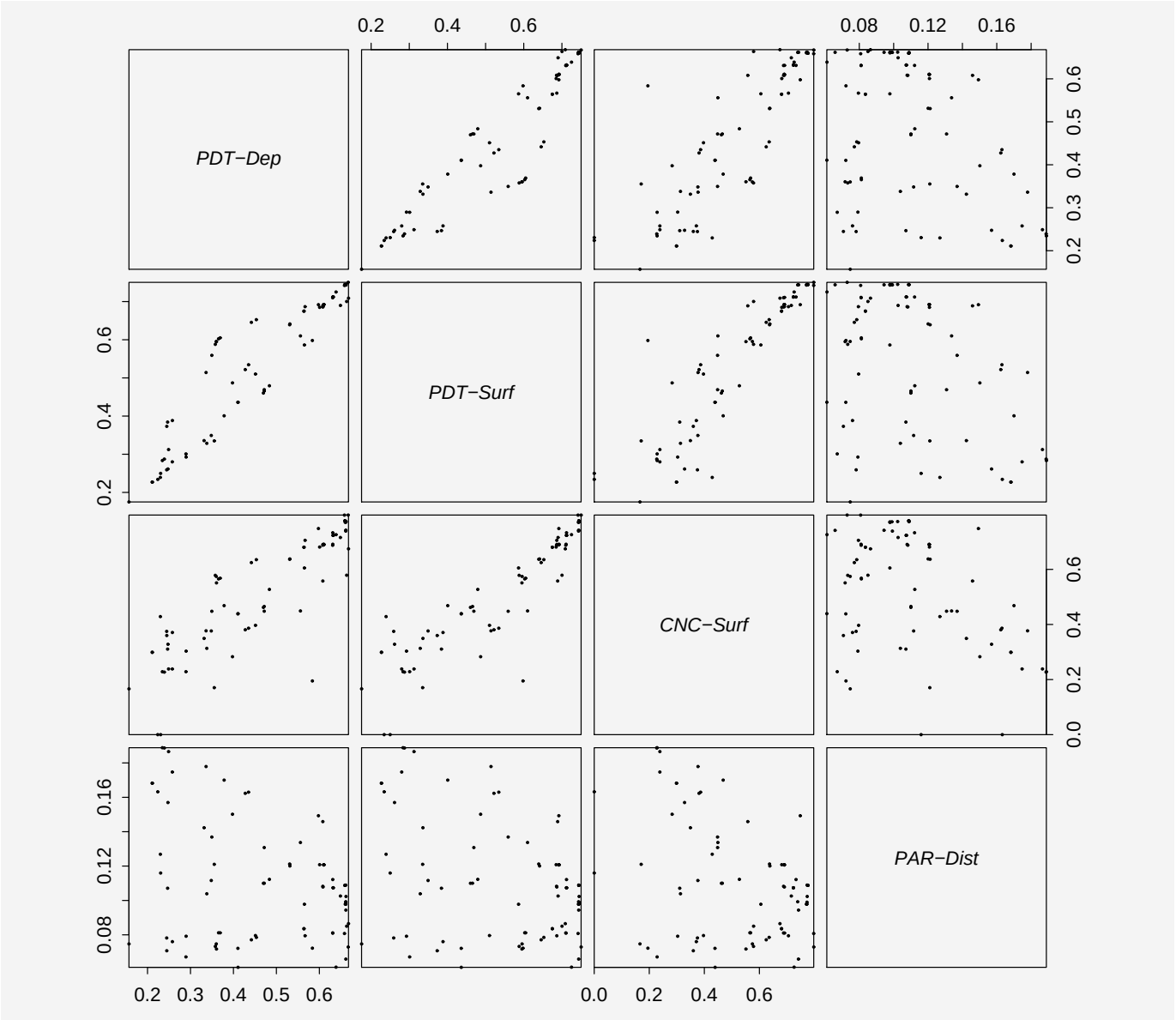
CNC-Surf



PAR-Dist



Comparison of AM evaluation results (MAP)



Extraction principle I

"Collocation components occur together more often than by chance"

- the corpus is interepreted as a sequence of randomly generated words
- word (*marginal*) probability ML estimations: $P(x) \approx \frac{f(x)}{N}$
- bigram (*joint*) probability ML estimations: $P(xy) \approx \frac{f(xy)}{N}$
- the chance $\sim$ the null hypothesis of independence: H_0: $\hat{P}(xy) = P(x) \cdot P(y)$

Example

Data: $f(\text{životní prostředí}) = 220$ *MLE:* $p(\text{životní prostředí}) = 0.000146$

$f(\text{životní}) = 261$ $p(\text{životní}) = 0.000174$

$f(\text{prostředí}) = 439$ $p(\text{prostředí}) = 0.000292$

H_0: $\hat{p}(\text{životní prostředí}) = p(\text{životní}) \cdot p(\text{prostředí}) = 0.0000000005$

$\hat{f}(\text{životní prostředí}) = 0.076$

AM: $PMI(\text{životní prostředí}) = \log \frac{f(\text{životní prostředí})}{\hat{f}(\text{životní prostředí})} = \log \frac{220}{0.076} = 11.53$

Extraction principle II

"Collocations occur as units in information-theoretically noisy environment"

- the corpus again interpreted as a sequence of randomly generated words
- at each point of the sequence we:
  1. estimate probability distribution of words occurring after (and before)
  2. measure uncertainty (*entropy*) what the next (or previous) word will be
- points with high uncertainty are likely to be collocation boundaries
- points with low uncertainty are likely to be located within a collocation

Example

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

Extraction principle III

“Collocations occur in different contexts to their components”

- non-compositionality: meaning of a collocation must differ from the *union* of the meaning of its components
- modeling meanings by empirical contexts: a bag of words occurring within a specified context window of a word or an expression
- the more different the contexts of an expression to its components are, the higher the chance is that the expression is a collocation

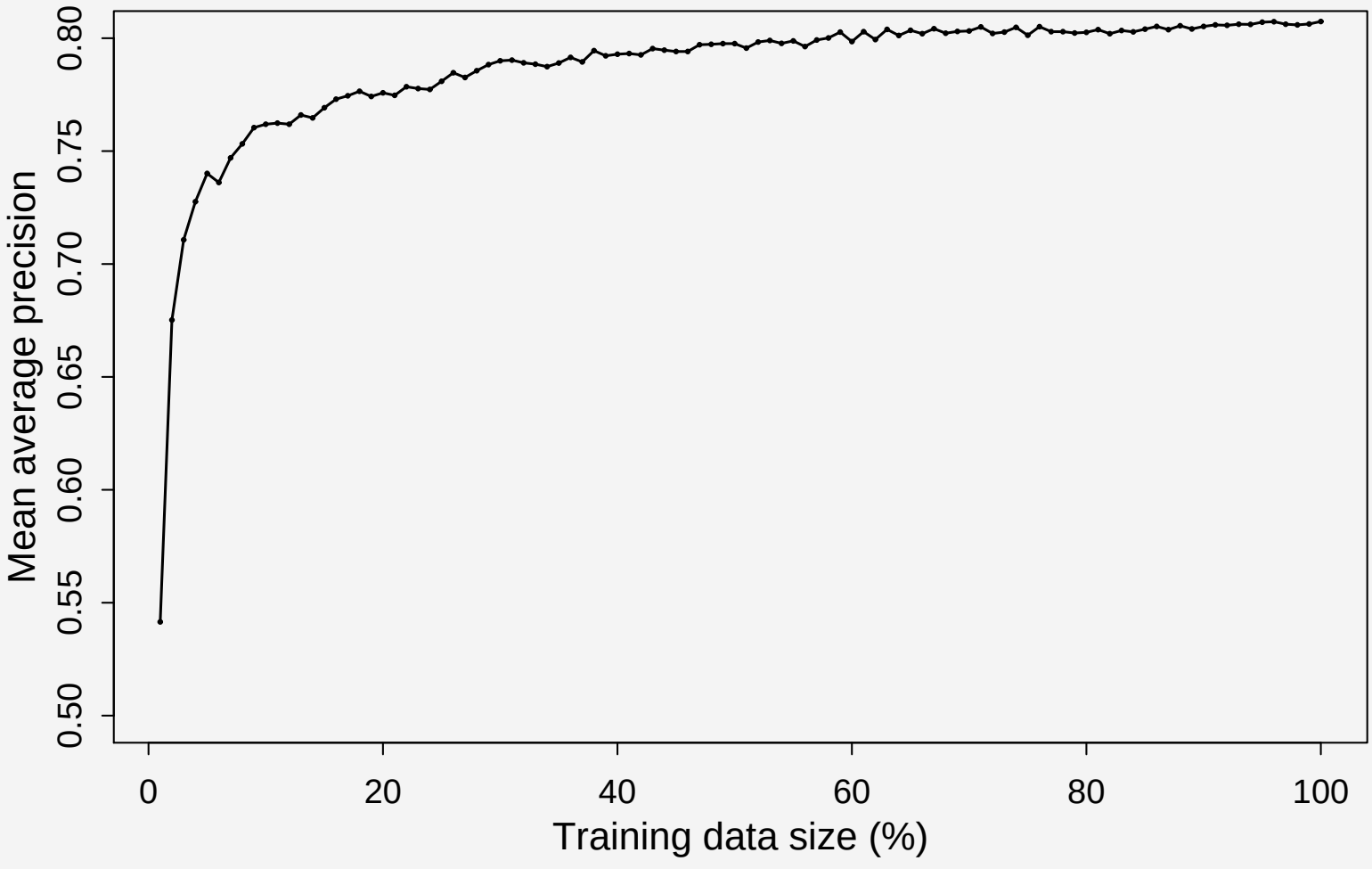
Example

. . . není. Maltské liry lze nakoupit pouze ve směnárnách, **černý trh** s valutami neexistuje. Na Maltě je v porovnání s . . .
. . . přestal. V patách za krizí vstoupil do Bělehradu **černý trh** , pašování a zvýšená kriminalita. Překupníci provážejí . . .
. . . nebyli z toho obviněni. Řídí gangy, které kontrolují **černý trh** a okrádají cizince. Oba byli zbaveni funkcí a byl . . .
. . . antidrogové hysterii. Následkem toho neexistoval ani **černý trh** , protože nebylo na čem vydělávat. V roce 1957 bylo . . .
. . . doručeny k rychlému zpracování. Naplno se již rozjíždí **černý trh** se vstupenkami. Na závod na 5000 m v rychlobruslařů . . .
. . . na čelném místě obchodu se zbraněmi. Zatímco **černý trh** se zbraněmi se pro celý svět stává čím dál tím větší. . . .
. . . čtením v parlamentu. Věřím, že brzy bude regulovat **černý trh** s ohroženými druhy zvířat, míní. Promoravské strany . . .

. . . jako malí čtyřletí a pětiletí kluci. Byl to dobytčí **trh** jako z minulého století. Se vším všudy prodávali . . .
. . . přání než reálných možností. Na rozdíl od dolaru se **trh** amerických státních dluhopisů nezměnil. A novými . . .
. . . opětnému nárůstu. Podle Plan Econu si český kapitálový **trh** bude v nejbližším roce počínat o něco lépe. Většina . . .
. . . To by mohlo vzhledem k propojení přes mezibankovní **trh** depozit vést k řetězovým reakcím. Příliv kapitálu . . .
. . . PVT, na ceně ztratil také indexový Tabák. Volný **trh** má však naštěstí i světlé stránky. K nim patří například . . .
. . . spoluzakladatel. Také v Maďarsku se uvolní mediální **trh** již letos. Maďarsko jako první z postkomunistických . . .
. . . . Mezi ně patří i OfficePorte Voice, který byl na **trh** uveden pod heslem ”více než modem”. Obsahuje totiž . . .

Learning curve

PDT-Dep



List of relevant publications

- Pavel Pecina: **Lexical Association Measures and Collocation Extraction**, *Multiword expressions: Hard going or plain sailing? Special issue of the International Journal of Language Resources and Evaluation*, Springer, 2009 (accepted).
- Pavel Pecina: **Machine Learning Approach to Multiword Expression Extraction**, *In Proceedings of the sixth International Conference on Language Resources and Evaluation (LREC) Workshop: Towards a Shared Task for Multiword Expressions*, Marrakech, Morocco, 2008.
- Pavel Pecina: **Reference Data for Czech Collocation Extraction**, *In Proceedings of the sixth International Conference on Language Resources and Evaluation (LREC) Workshop: Towards a Shared Task for Multiword Expressions*, Marrakech, Morocco, 2008.
- Pavel Pecina, Pavel Schlesinger: **Combining Association Measures for Collocation Extraction**, *In Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, Sydney, Australia, 2006.
- Silvie Cinková, Petr Podveský, Pavel Pecina, Pavel Schlesinger: **Semi-automatic Building of Swedish Collocation Lexicon**, *In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genova, Italy, 2006.
- Pavel Pecina: **An Extensive Empirical Study of Collocation Extraction Methods**, *In Proceedings of the Association for Computational Linguistics Student Research Workshop (ACL)*, Ann Arbor, Michigan, USA, 2005.
- Pavel Pecina, Martin Holub: **Semantically Significant Collocations**, *UFAL/CKL Technical Report TR-2002-13*, Faculty of Mathematics and Physics, Charles University, Prague, Czech Rep., 2002.