Combining Association Measures for Collocation Extraction

Pavel Pecina & **Pavel Schlesinger**

Institute of Formal and Applied Linguistics
Charles University, Prague, Czech Republic

{pecina,schlesinger}@ufal.mff.cuni.cz

Outline

- Notion of collocation
  … *with definitions, characteristic properties, and examples.*
- Manual annotation of reference data
  … *of reasonable size and quality.*
- Collocation extraction and evaluation of basic methods
  … *by means of precision-recall curves and mean average precision.*
- Combining association measures for collocation extraction
  … *to improve results of individual measures.*
- Reduce number of combined measures
  … *to remove redundant and ineffective measures.*

Notion of Collocation

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."

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."

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 → valency
b) extensionally" → collocations

Characteristic Properties of Collocations

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)*

Reference Data Annotation

Source: Prague Dependency Treebank 2.0 *(new version of the Czech treebank)*
Sentences: 87,980 *(morphological and analytical annotation)*
Word forms: 1,504,847
Dependency bigram types: 635,952
Reference bigram types (f>5): 26,450 *(frequency filtering)*
Reference collocation candidates: 12,232 *(Part-of-Speech filtering)*

Focus on bigram collocations:

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

Manual annotation:

- The list of collocation candidates processed by three linguists in paralel.
- Bigrams that all three annotators independently recognized as collocations (of any type) were considered true collocations (20.9%).

Collocation categories and types

- idiomatic expressions
  studená válka (cold war), *natáhnout bačkory (kick the bucket)*
- technical terms
  předseda vlády (prime minister), *očitý svědek (eye witness)*
- support verb constructions
  mít pravdu (to be right), *učinit rozhodnutí (make decision)*
- names of persons, locations, and other entities
  Pražský hrad (Prague Castle), *Červený kříž (Red Cross)*
- stock phrases
  zásadní problém (major problem), *konec roku (end of the year)*

Colllocation 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 and can be used for ranking (highest to the top) or classification (by setting a treshold).

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* ← *threshold*
system type	*0*
and others	*0*
program in	*0*
level is	*0*

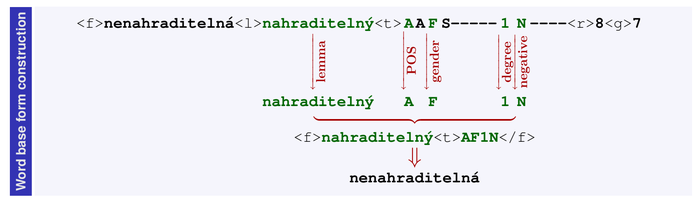
Methodology

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
   - empirical contexts
4. Computing association measures
5. Ranking or classification

Word Base Forms

- Surface word forms too specific (*rich Czech morphology*)
- Pure lemmas too general (*loss of syntactic and semantic information*)
- Lemmas with a subset of morphological tag fairly optimal

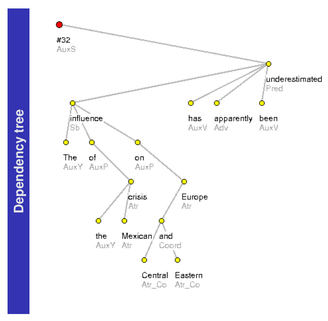
Word base form construction



Dependency Bigrams

- Dependency trees broken down to dependency bigrams consisting of word base forms, dependency type and direction:

Dependency tree



Part-of-speech

A N	*black market*
N N	*crown prince*
V N	*give birth*
D A	*mentally retarded*

Dependency

Atr	*strong tea*
Sb	*prices drop*
Obj	*commit crime*
Adv	*incredibly beautiful*

Direction

velký výr (big owl)
výr velký (owl big ≈ Eagle Owl)

Coocurrence and Context Statistics

- Contingency tables – observed frequencies and marginal frequencies:

Contingency table

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

Example

	X=*black*	$X\neq$*black*	
Y=*market*	*black market*	*new market*	$*$ *market*
$Y\neq$*market*	*black horse*	*new horse*	$*$ *horse*
	black $*$	*new* $*$	$*$ $*$

- Immediate context – words immediately preceding or following the bigram:

Left context

…součástí trhu, vznikl **obratem černý trh** s plyšovými medvídky a …
…zabránit přísunu drog na **domácí černý trh** v hodnotě 32 milionu …
…stejnými jednotlivci i **kompletní černý trh** . Jinými slovy, byla by …
…pomáhali pašování cigaret na **černý trh** do východního Německa. …
…nájemních práv **nezaručený černý trh** . Libor Dellin, člen …
…pašovaného zboží a **kvůli černý trh** jsou toho výmluvným …
…Také například tím, že **vznikne černý trh** , který je ke spotřebitelům …
…nabídku a pak **nastupuje černý trh** . Za možnost přestupu na …
…Řídí gangy, které **kontrolují černý trh** a okrádají cizince. Oba …
…nájemného " bylo a je **omezit černý trh** s byty, nestane se nic. …

Right context

…nějak negativně tento **černý trh naše** hospodářství? Je to …
…inzeráty. Rozmohl se **černý trh bytů** a skutečné náklady na …
…jak se říká na Arbatu, **černý trh něco** do sebe. Je - li hlad …
…Naplno se již rozjíždí **černý trh se** vstupenkami. Na závod …
…staročitnostmi měl řídit, **černý trh podporuje** na straně jedné …
…Našim lidem pro samý **černý trh nezbývají** čas na sex, a tak …
…unie však ukazují, že **černý trh překonal** stagnaci a pirátí…
…ceny, funguje čilý **černý trh dosud** . Země bez chudých …
…novými zbraněmi. Na **černý trh odhalený** specialisty z úřadu …
…se vlastně jedná o **černý trh s** byty. Přípustí ovšem, že …

- Empirical context – words occuring within a specified context window:

Bigram context

…180 miliónů dolarů. **Ovlivňuje nějak negativně tento černý trh naše** hospodářství? Je to pouze ztráta na daních …
…není. **Maltské liry lze nakoupit pouze ve směnárnách, černý trh s** valutami neexistuje. Na Maltě je v porovnání s …
…operoval ženu. **A přece má, jak se říká na Arbatu, černý trh něco do** sebe. Je - li hlad nejlepší kuchař, je …
…přestal. **V patách za krizí vstoupil do Bělehradu černý trh , pašování a zvýšená kriminalita.** Překupníci provážejí …
…nebyli z toho obvinění. **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 celném 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.** mlní. Promoravské strany …
…100 tisíc korun. **Podle Přihoe se vlastně jedná o černý trh s byty.** Přípustil ovšem, že právě v případě bytového…

Word context

…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 …
…velkými těžkostmi. **Ale dostáváme se do situace, že trh nechce ani je.** Měli bychom směřovat k tomu, abychom …
…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 …
…pozastavena. **Podle pracovníků závodu se na český trh nedostalo ani kilo.** Zda je maso z Belgie závadné …
…korun. **To, že jsme si tyto obchodníky pustili na náš trh , bylo velmi krátkozraké.** Likvidujeme své vlastní …

Collocation Hypotheses and Types of Association Measures

$\mathbf{H}_1$: *"Collocations are very frequent word combinations."*

- ML estimations of joint and conditional probabilities

$\mathbf{H}_2$: *"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

$\mathbf{H}_3$: *"Collocations occur as units in a (inf.-theoretically) noisy environment."*

- Information-theory measures of immediate contexts

$\mathbf{H}_4$: *"Collocations occur in different contexts than their components."*

- Information-theory context measures
- Information-retrieval context similarity measures

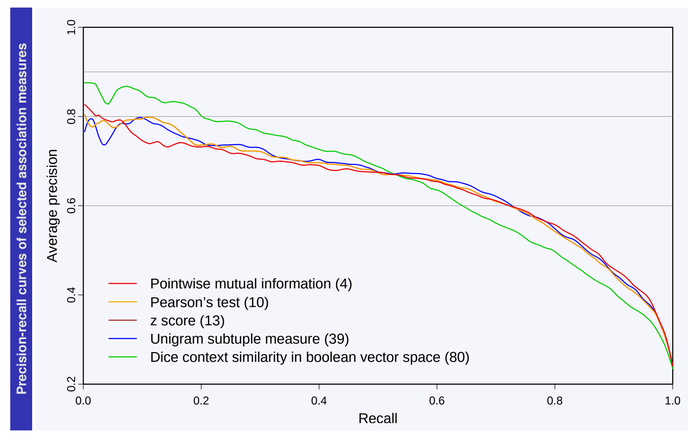
Association Measures

1. *Joint probability*	$P(xy)$
*2. *Conditional probability*	$P(y\|x)$
3. *Reverse conditional prob.*	$P(x\|y)$
4. *Pointwise mutual inform.*	$\log\frac{P(xy)}{P(x*)P(*y)}$
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*	$\sum_{ij}\frac{(f_{ij}-\hat{f}_{ij})^2}{\hat{f}_{ij}}$
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*	$\frac{f(xy)-\hat{f}(xy)}{\sqrt{\hat{f}(xy)(1-(f(xy)/N))}}$
14. *Poison significance meas.*	$\frac{f(xy)-\hat{f}(xy)\log f(xy)+\log f(xy)!}{\log N}$
15. *Log likelihood ratio*	$-2\sum_{ij}f_{ij}\log\frac{f_{ij}}{\hat{f}_{ij}}$
16. *Squared log likelihood rat.*	$-2\sum_{ij}\frac{\log f_{ij}^2}{\hat{f}_{ij}}$
Association coefficients:	
17. *Russel-Rao*	$\frac{a}{a+b+c+d}$
18. *Sokal-Michiner*	$\frac{a+d}{a+b+c+d}$
19. *Rogers-Tanimoto*	$\frac{a+d}{a+2b+2c+d}$
20. *Hamann*	$\frac{(a+d)-(b+c)}{a+b+c+d}$
21. *Third Sokal-Sneath*	$\frac{b+c}{a+d}$
22. *Jaccard*	$\frac{a}{a+b+c}$
*23. *First Kulczynsky*	$\frac{a}{b+c}$
24. *Second Sokal-Sneath*	$\frac{a}{a+2(b+c)}$
25. *Second Kulczynski*	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
*26. *Fourth Sokal-Sneath*	$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{d+b}+\frac{d}{d+c}\right)$
*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*	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
46. ***J* measure**	$\max[P(xy)\log\frac{P(x\|y)}{P(*y)}+P(x\bar{y})\log\frac{P(\bar{x}\|y)}{P(*\bar{y})}, P(xy)\log\frac{P(x\|y)}{P(x*)}+P(\bar{x}y)\log\frac{P(\bar{x}\|y)}{P(\bar{x}*)}]$
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*	$\max[\frac{P(x*)P(*y)}{P(x\bar{y})},\frac{P(x*)P(*y)}{P(\bar{x}y)}]$
51. *Piatersky-Shapiro*	$P(xy)-P(x*)P(*y)$
52. *Certainity factor*	$\max[\frac{P(y\|x)-P(*y)}{1-P(*y)},\frac{P(x\|y)-P(x*)}{1-P(x*)}]$
53. *Added value (AV)*	$\max[P(y\|x)-P(*y),P(x\|y)-P(x*)]$
54. *Collective strength*	$\frac{P(xy)+P(\bar{x}\bar{y})}{P(x*)P(y)+P(\bar{x}*)P(*y)}\cdot\frac{1-P(x*)P(*y)-P(\bar{x}*)P(*y)}{1-P(xy)-P(\bar{x}\bar{y})}$
*55. *Klosgen*	$\sqrt{P(xy)}\cdot AV$
Context measures:	
*56. *Context entropy*	$-\sum_w P(w\|C_{xy})\log P(w\|C_{xy})$
*57. *Left context entropy*	$-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
58. *Right context entropy*	$-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
59. *Left context divergence*	$P(x*)\log P(x*)-\sum_w P(w\|C^l_{xy})\log P(w\|C^l_{xy})$
60. *Right context divergence*	$P(*y)\log P(*y)-\sum_w P(w\|C^r_{xy})\log P(w\|C^r_{xy})$
61. *Cross entropy*	$-\sum_w P(w\|C_x)\log P(w\|C_y)$
62. *Reverse cross entropy*	$-\sum_w P(w\|C_y)\log P(w\|C_x)$
63. *Intersection measure*	$\frac{2\|C_x\cap C_y\|}{\|C_x\|+\|C_y\|}$
*64. *Euclidean norm*	$\sqrt{\sum_w(P(w\|C_x)-P(w\|C_y))^2}$
65. *Cosine norm*	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2\cdot\sum_w P(w\|C_y)^2}$
*66. ***L1* norm**	$\sum_w\|P(w\|C_x)-P(w\|C_y)\|$
67. *Confusion probability*	$\sum_w\frac{P(x\|C_w)P(y\|C_w)P(w)}{P(x*)}$
*68. *Reverse confusion prob.*	$\sum_w\frac{P(y\|C_w)P(x\|C_w)P(w)}{P(*y)}$
*69. *Jensen-Shannon diverg.*	$\frac{1}{2}[D(p(w\|C_x)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))+D(p(w\|C_y)\|\|\frac{1}{2}(p(w\|C_x)+p(w\|C_y)))]$
*70. *Cosine of pointwise* ***MI***	$\frac{\sum_w MI(ux)MI(uy)}{\sqrt{\sum_w MI(ux)^2}\cdot\sqrt{\sum_w MI(uy)^2}}$
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*	$D(p(w\|C_x)\|\|\alpha(w\|C_y)+(1-\alpha)p(w\|C_x))$
74. *Reverse skew divergence*	$D(p(w\|C_y)\|\|\alpha p(w\|C_x)+(1-\alpha)p(w\|C_y))$
75. *Phrase word coocurrence*	$\frac{1}{2}\left(\frac{f(x\bar{C}_{xy})}{f(x*)}+\frac{f(\bar{y}C_{xy})}{f(*y)}\right)$
76. *Word association*	$\frac{1}{2}\left(\frac{f(x\bar{C}_y)-f(xy)}{f(x*)}+\frac{f(\bar{y}C_x)-f(xy)}{f(*y)}\right)$
Cosine context similarity:	$\frac{1}{2}(\cos(\mathbf{c}_x,\mathbf{c}_{xy})+\cos(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \cos(\mathbf{c}_x,\mathbf{c}_y)=\frac{\sum x_iy_i}{\sqrt{\sum x_i^2}\cdot\sqrt{\sum y_i^2}}$
*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·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:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_iy_i}{\sum x_i^2+\sum y_i^2}$
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·idf vector space*	$z_i=f(w_i\|C_z)\cdot\frac{N}{df(w_i)};\ df(w_i)=\|\{x:w_i\in C_x\}\|$

Evaluation of Assosciation Measures

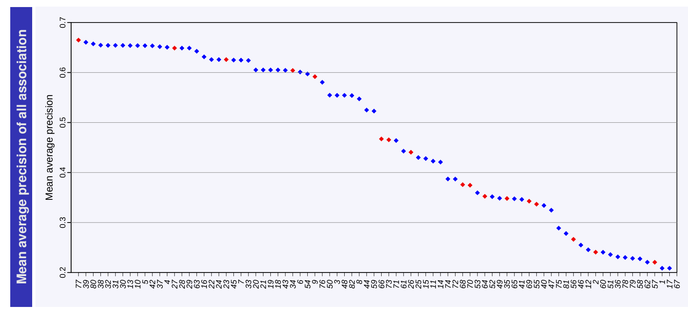
- Precision: fraction of positive predictions correct (*given a threshold*)
- Recall: fraction of positives correctly predicted (*given a threshold*)
- Precision–recall curve: PR scores for all possible threshold values

Precision-recall curves of selected association measures



- Average precision: expected value of precision for all possible values of recall
- Mean average precision: mean of AP computed for crossvalidation data folds

Mean average precision of all asssociation

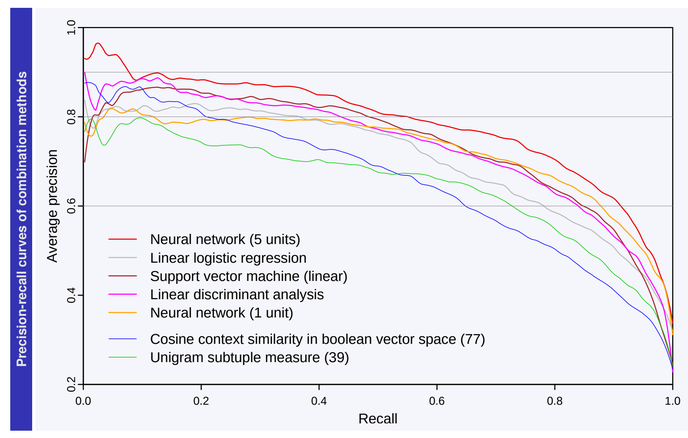


Combining Association Measures

- Each collocation candidate x^i is described by the feature vector $\mathbf{x}^i=(x^i_1,\ldots,x^i_{82})^T$ consisting of scores of all association measures.
- We look for a ranker function $f(\mathbf{x})\rightarrow\mathbb{R}$ that determines the strength of lexical association between components of bigram x:

- Linear logistic regression — $\mathrm{logit}\,(\pi)=\beta_0+\beta_1x_1+\ldots+\beta_px_p$
- Linear discriminant analysis — $\max_{\mathbf{c}}\mathbf{c}^T\boldsymbol{B}\mathbf{c}/\mathbf{c}^T\boldsymbol{W}\mathbf{c}$
- Support vector machines — $\min_{\beta_0,\boldsymbol{\beta}}\sum_{i=1}^n\left[1-y^i(\beta_0+\boldsymbol{\beta}^T\mathbf{x}^i)\right]^+_+\frac{\lambda}{2}||\boldsymbol{\beta}||^2$
- Neural networks — $y^i=\phi_0\left(\alpha_0+\sum w_{hi}\phi_h(\alpha_h+\sum w_{jh}x_j)\right)$

Precision-recall curves of combination methods



Mean average precision of combination methods

method	*AP*			*MAP*	
	R=20	*R=50*	*R=80*	*R=(0.1,0.9)*	+
Neural Network (5 units)	89.56	82.74	70.11	80.81	21.53
Neural Network (3 units)	89.41	81.99	69.64	79.71	19.88
Neural Network (2 units)	86.92	81.68	68.33	78.77	18.47
Support Vector Machine (linear)	85.72	79.49	63.86	75.66	13.79
Linear Discriminant Analysis	84.72	77.18	62.90	75.11	12.96
Support Vector Machine (quadratic)	84.29	79.54	64.24	74.53	12.09
Neural Network (1 unit)	77.98	76.83	66.75	73.25	10.17
Logistic Linear Regression	82.45	76.26	58.61	71.88	8.11
Cosine similarity (77)	80.94	68.90	50.54	66.49	0.00
Unigram subtuples (39)	74.55	67.49	55.16	65.74	-

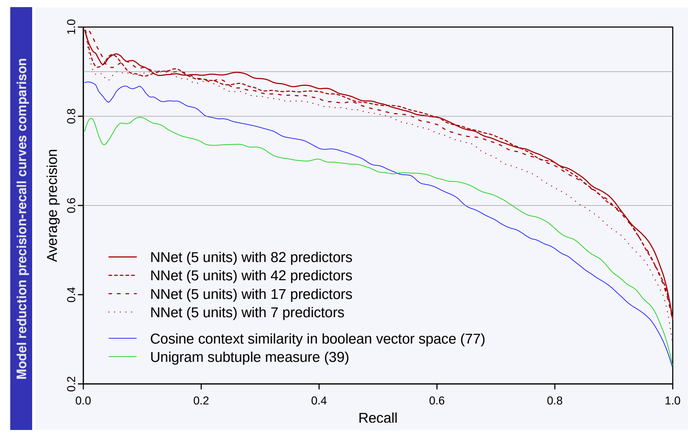
Model Reduction

- Combination models are too complex in number of predictors used.
- Some association measures are very similar (*analytically or empirically*) and as predictors perhaps even redundant. Such measures have no use in the models, make their training problematic, and should be excluded.

The model reduction algorithm:

1. Starts with hierarchical clustering of variables in order to group those with similar contribution to the model (by Pearson's correlation coefficient).
2. After $82-d$ iteration, the variables are grouped into d clusters, one representative from each cluster is selected as a predictor into the initial model.
3. In each next step, the algorithm removes one predictor causing minimal degradation of performance (measured by MAP on held–out data).

Model reduction precision-recall curves comparison



Results

- The best performing individual association measures are cosine context similarity in boolean vector space and unigram subtuple measure.
- The best performing combination method is Neural Network with 5 units in the hidden layer with 21.53% relative improvement in terms of MAP.
- Number of variables in this model was reduced to 17 without significant degradation of its performance.

Coling/ACL, 17-21 July 2006, Sydney, Australia