Lexical Association Measures

Collocation Extraction

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Charles University in Prague

LOEWE Digital Humanities
Goethe University Frankfurt am Main
12.7.2012

Talk outline

Lexical association

Semantic association

- reflects semantic relationship between words
- *synonymy, antonymy, hyponymy, meronymy, etc.* → stored in a thesaurus

  sick – ill, baby – infant, dog – cat

Cross-language association

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

  maison$_{(FR)}$ – *house*$_{(EN)}$, *baum*$_{(GE)}$ – *tree*$_{(EN)}$, *květina*$_{(CZ)}$ – *flower*$_{(EN)}$

Collocational association

- restricts combination of words into phrases (beyond grammar!)
- *collocations / multiword expressions* → stored in a lexicon

  crystal clear, cosmetic surgery, cold war

Lexical association

Semantic association

- reflects semantic relationship between words
- *synonymy, antonymy, hyponymy, meronymy, etc.* → stored in a thesaurus

  sick – ill, baby – infant, dog – cat

Cross-language association

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

  maison$_{(FR)}$ *– house*$_{(EN)}$, *baum*$_{(GE)}$ *– tree*$_{(EN)}$, *květina*$_{(CZ)}$ *– flower*$_{(EN)}$

Collocational association

- restricts combination of words into phrases (beyond grammar!)
- *collocations / multiword expressions* → stored in a lexicon

  crystal clear, cosmetic surgery, cold war

Lexical association

Semantic association

- reflects semantic relationship between words
- *synonymy, antonymy, hyponymy, meronymy, etc.* → stored in a thesaurus

  sick – ill, baby – infant, dog – cat

Cross-language association

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

  maison$_{(FR)}$ – house$_{(EN)}$, baum$_{(GE)}$ – tree$_{(EN)}$, květina$_{(CZ)}$ – flower$_{(EN)}$

Collocational association

- restricts combination of words into phrases (beyond grammar!)
- *collocations / multiword expressions* → stored in a lexicon

  crystal clear, cosmetic surgery, cold war

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
- collocation extraction

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
- collocation extraction

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
- 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 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 advance the state of the art in collocation extraction

Limitations

- ✓ focus on bigram (*two-word*) collocations
  (limited scalability to higher-order n-grams; limited corpus size)
- ✓ binary (*two-class*) discrimination only (*collocation/non-collocation*)

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 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 advance the state of the art in collocation extraction

Limitations

- ✓ focus on bigram (*two-word*) collocations
  (limited scalability to higher-order n-grams; limited corpus size)
- ✓ binary (*two-class*) discrimination only (*collocation/non-collocation*)

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 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 advance the state of the art in collocation extraction

Limitations

✓ focus on bigram (*two-word*) collocations
(limited scalability to higher-order n-grams; limited corpus size)

✓ binary (*two-class*) discrimination only (*collocation/non-collocation*)

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*

  **Colorless green ideas sleep furiously (N. Chomsky)*
- must be adhered to in order to produce correct, meaningful, fluent utterances
- ranges from almost free word combinations to very fixed word expressions
- specified intensionally: by general rules based on *common properties of words* or extensionally: by specific constraints for *particular words*

Collocations

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

Types of collocations

idioms *(to kick the bucket)*, proper names *(New York)*, technical terms *(hard disk)*, phrasal verbs *(to look after)*, light verb compounds *(to take a nap)*, lexically restricted expressions *(broad daylight)*, etc.

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*

  **Colorless green ideas sleep furiously (N. Chomsky)*
- must be adhered to in order to produce correct, meaningful, fluent utterances
- ranges from almost free word combinations to very fixed word expressions
- specified intensionally: by general rules based on *common properties of words* or extensionally: by specific constraints for *particular words*

Collocations

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

Types of collocations

idioms *(to kick the bucket)*, proper names *(New York)*, technical terms *(hard disk)*, phrasal verbs *(to look after)*, light verb compounds *(to take a nap)*, lexically restricted expressions *(broad daylight)*, etc.

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*

  **Colorless green ideas sleep furiously (N. Chomsky)*
- must be adhered to in order to produce correct, meaningful, fluent utterances
- ranges from almost free word combinations to very fixed word expressions
- specified intensionally: by general rules based on *common properties of words* or extensionally: by specific constraints for *particular words*

Collocations

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

Types of collocations

idioms *(to kick the bucket)*, proper names *(New York)*, technical terms *(hard disk)*, phrasal verbs *(to look after)*, light verb compounds *(to take a nap)*, lexically restricted expressions *(broad daylight)*, etc.

Collocation properties

Semantic non-compositionality

- exact meaning cannot be (fully) inferred from the meaning of components
  to kick the bucket

Syntactic non-modifiability

- syntactic structure cannot be freely modified (*word order, word insertions etc.*)
  poor as a church mouse vs. *poor as a *fat church mouse*

Lexical non-substitutability

- components cannot be substituted by their synonyms or other words
  stiff breeze vs. **stiff wind*

Translatability into other languages

- translation cannot generally be performed blindly, in a word-by-word manner
  ice cream – zmrzlina

Domain dependency

- collocational character only in specific domains
  carriage return

Collocation properties

Semantic non-compositionality

- exact meaning cannot be (fully) inferred from the meaning of components
  to kick the bucket

Syntactic non-modifiability

- syntactic structure cannot be freely modified (*word order, word insertions etc.*)
  poor as a church mouse vs. *poor as a *fat church mouse*

Lexical non-substitutability

- components cannot be substituted by their synonyms or other words
  stiff breeze vs. **stiff wind*

Translatability into other languages

- translation cannot generally be performed blindly, in a word-by-word manner
  ice cream – zmrzlina

Domain dependency

- collocational character only in specific domains
  carriage return

Collocation properties

Semantic non-compositionality

- exact meaning cannot be (fully) inferred from the meaning of components
  to kick the bucket

Syntactic non-modifiability

- syntactic structure cannot be freely modified (*word order, word insertions etc.*)
  poor as a church mouse vs. *poor as a *fat church mouse*

Lexical non-substitutability

- components cannot be substituted by their synonyms or other words
  stiff breeze vs. **stiff wind*

Translatability into other languages

- translation cannot generally be performed blindly, in a word-by-word manner
  ice cream – zmrzlina

Domain dependency

- collocational character only in specific domains
  carriage return

Collocation properties

Semantic non-compositionality

- exact meaning cannot be (fully) inferred from the meaning of components
  to kick the bucket

Syntactic non-modifiability

- syntactic structure cannot be freely modified (*word order, word insertions etc.*)
  poor as a church mouse vs. *poor as a *fat church mouse*

Lexical non-substitutability

- components cannot be substituted by their synonyms or other words
  stiff breeze vs. **stiff wind*

Translatability into other languages

- translation cannot generally be performed blindly, in a word-by-word manner
  ice cream – zmrzlina

Domain dependency

- collocational character only in specific domains
  carriage return

Collocation properties

Semantic non-compositionality

- exact meaning cannot be (fully) inferred from the meaning of components
  to kick the bucket

Syntactic non-modifiability

- syntactic structure cannot be freely modified (*word order, word insertions etc.*)
  poor as a church mouse vs. *poor as a *fat church mouse*

Lexical non-substitutability

- components cannot be substituted by their synonyms or other words
  stiff breeze vs. **stiff wind*

Translatability into other languages

- translation cannot generally be performed blindly, in a word-by-word manner
  ice cream – zmrzlina

Domain dependency

- collocational character only in specific domains
  carriage return

Collocation extraction

Task

- to extract a list of collocations (*types*) from a text corpus
- 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"*

Collocation extraction

Task

- to extract a list of collocations (*types*) from a text corpus
- 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"*

Collocation extraction

Task

- to extract a list of collocations (*types*) from a text corpus
- 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 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: $p(x) = \frac{f(x)}{N}$
- bigram (*joint*) probability ML estimations: $p(xy) = \frac{f(xy)}{N}$
- the chance $\sim$ the null hypothesis of independence: H_0: $\hat{p}(xy) = p(x) \cdot p(y)$

AM: *Log-likelihood ratio, χ^2 test, Odds ratio, Jaccard, Pointwise mutual information*

Example: *Pointwise Mutual Information*

Data: $f(\textit{iron curtain}) = 11$ MLE: $p(\textit{iron curtain}) = 0.000007$

$f(\textit{iron}) = 30$ $p(\textit{iron}) = 0.000020$

$f(\textit{curtain}) = 15$ $p(\textit{curtain}) = 0.000010$

H_0: $\hat{p}(\textit{iron curtain}) = p(\textit{iron}) \cdot p(\textit{curtain}) = 0.000000000020$

AM: $PMI(\textit{iron curtain}) = log\frac{p(xy)}{\hat{p}(xy)} = log\frac{0.000007}{0.000000000020} = 18.417$

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: $p(x) = \frac{f(x)}{N}$
- bigram (*joint*) probability ML estimations: $p(xy) = \frac{f(xy)}{N}$
- the chance $\sim$ the null hypothesis of independence: H_0: $\hat{p}(xy) = p(x) \cdot p(y)$

AM: *Log-likelihood ratio, χ^2 test, Odds ratio, Jaccard, Pointwise mutual information*

Example: *Pointwise Mutual Information*

Data: $f(\textit{iron curtain}) = 11$, $f(\textit{iron}) = 30$, $f(\textit{curtain}) = 15$

MLE: $p(\textit{iron curtain}) = 0.000007$, $p(\textit{iron}) = 0.000020$, $p(\textit{curtain}) = 0.000010$

H_0: $\hat{p}(\textit{iron curtain}) = p(\textit{iron}) \cdot p(\textit{curtain}) = 0.000000000020$

AM: $PMI(\textit{iron curtain}) = log\frac{p(xy)}{\hat{p}(xy)} = log\frac{0.000007}{0.000000000020} = 18.417$

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 estimate:
  1. probability distribution of words occurring after/before: $\mathbf{p}(w|C^r_{xy})$, $\mathbf{p}(w|C^l_{xy})$
  2. uncertainty (entropy) what the next/previous word is: $H(\mathbf{p}(w|C^r_{xy}))$,$H(\mathbf{p}(w|C^l_{xy}))$
- points with high uncertainty are likely to be collocation boundaries
- points with low uncertainty are likely to be located within a collocation

AM: *Left context entropy, Right context entropy*

Example: $H(\mathbf{p}(w|C^r_{xy}))$

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

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 estimate:
  1. probability distribution of words occurring after/before: $\mathbf{p}(w|C^r_{xy})$, $\mathbf{p}(w|C^l_{xy})$
  2. uncertainty (entropy) what the next/previous word is: $H(\mathbf{p}(w|C^r_{xy}))$,$H(\mathbf{p}(w|C^l_{xy}))$
- points with high uncertainty are likely to be collocation boundaries
- points with low uncertainty are likely to be located within a collocation

AM: *Left context entropy, Right context entropy*

Example: $H(\mathbf{p}(w|C^r_{xy}))$

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

AM: *J-S divergence, K-L divergence, Skew divergence, Cosine similarity in vector space*

Example: C_{xy}, C_x

…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 …

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

AM: *J-S divergence, K-L divergence, Skew divergence, Cosine similarity in vector space*

Example: C_{xy}, C_x

. . . 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 . . .

Inventory of 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**	$\frac{P(xy)+P(\bar{x}\bar{y})-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}{1-P(x*)P(*y)-P(\bar{x}*)P(*\bar{y})}$
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**	$\frac{\sum_w P(w\|C_x)P(w\|C_y)}{\sum_w P(w\|C_x)^2 \cdot \sum_w P(w\|C_y)^2}$
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* · *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:	$\frac{1}{2}(\mathrm{dice}(\mathbf{c}_x,\mathbf{c}_{xy})+\mathrm{dice}(\mathbf{c}_y,\mathbf{c}_{xy}))$ $\mathbf{c}_z=(z_i);\ \mathrm{dice}(\mathbf{c}_x,\mathbf{c}_y)=\frac{2\sum x_i y_i}{\sum x_i{}^2+\sum y_i{}^2}$
80.	**in boolean vector space**	$z_i=\delta(f(w_i\|C_z))$
81.	**in *tf* vector space**	$z_i=f(w_i\|C_z)$
82.	**in *tf* · *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.

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*)
5. application of a choosen lexical association measure to all candidates
6. ranking/classification of collocation candidates according to their scores

Ranking

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

Classification

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

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*)
5. application of a choosen lexical association measure to all candidates
6. ranking/classification of collocation candidates according to their scores

Ranking

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

Classification

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

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*)
5. application of a choosen lexical association measure to all candidates
6. ranking/classification of collocation candidates according to their scores

Ranking

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

Classification

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

Reference data set

Source corpus

- Prague Dependency Treebank 2.0, 1.5 mil. tokens
- manually annotated with *morphological* and *syntactic dependency* information

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*)
- 6 categories, merged into two: collocations (1-5), non-collocations (0):

  5. *idiomatic expressions*
  4. *technical terms*
  3. *support verb constructions*
  2. *proper names*
  1. *frequent unpredictable usages*
  0. *non-collocations*

- 12 232 candidates = 2 557 true collocations + 9 675 true non-collocations

Reference data set

Source corpus

- Prague Dependency Treebank 2.0, 1.5 mil. tokens
- manually annotated with *morphological* and *syntactic dependency* information

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*)
- 6 categories, merged into two: collocations (1-5), non-collocations (0):

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

---

0. *non-collocations*

- 12 232 candidates = 2 557 true collocations + 9 675 true non-collocations

Reference data set

Source corpus

- Prague Dependency Treebank 2.0, 1.5 mil. tokens
- manually annotated with *morphological* and *syntactic dependency* information

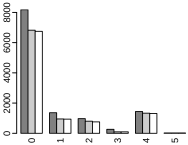
Collocation candidates

- 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*)
- 6 categories, merged into two: collocations (1-5), non-collocations (0):

5. *idiomatic expressions*
4. *technical terms*
3. *support verb constructions*
2. *proper names*
1. *frequent unpredictable usages*
0. *non-collocations*



- 12 232 candidates = 2 557 true collocations + 9 675 true non-collocations

Experimental design

Reference data

- split into 7 stratified folds of the same size (the same ratio of true collocations)
- 1 fold put aside as held-out data
- 6 folds used for evaluation of AMs

eval$_1$	***eval$_2$***	***eval$_3$***	***eval$_4$***	***eval$_5$***	***eval$_6$***	***held-out***

Evaluation

- based on quality of ranking (*ranking performance*)
- evaluation measures estimated on each eval fold separately and averaged

Significance testing

- methods compared by paired Wilcoxon signed-ranked test on the 6 eval folds
- significance level $\alpha = 0.05$

Experimental design

Reference data

- split into 7 stratified folds of the same size (the same ratio of true collocations)
- 1 fold put aside as held-out data
- 6 folds used for evaluation of AMs

$eval_1$	$eval_2$	$eval_3$	$eval_4$	$eval_5$	$eval_6$	held-out

Evaluation

- based on quality of ranking (*ranking performance*)
- evaluation measures estimated on each eval fold separately and averaged

Significance testing

- methods compared by paired Wilcoxon signed-ranked test on the 6 eval folds
- significance level $\alpha = 0.05$

Experimental design

Reference data

- split into 7 stratified folds of the same size (the same ratio of true collocations)
- 1 fold put aside as held-out data
- 6 folds used for evaluation of AMs

$eval_1$	$eval_2$	$eval_3$	$eval_4$	$eval_5$	$eval_6$	held-out

Evaluation

- based on quality of ranking (*ranking performance*)
- evaluation measures estimated on each eval fold separately and averaged

Significance testing

- methods compared by paired Wilcoxon signed-ranked test on the 6 eval folds
- significance level $\alpha = 0.05$

Evaluation measures: Precision – Recall

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

Evaluation measures: Precision – Recall

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

Ranking	
red cross	*15.66*
iron curtain	*15.23*
decimal point	*14.01*
coupon book	*13.83*
book author	*11.05*
arithmetic operation	*10.52*
paper feeder	*10.17*
new book	*10.09*
round table	*7.03*
new wave	*6.59*
gas station	*6.04*
system type	*3.54*
central part	*1.54*
and others	*0.54*
program in	*0.35*
level is	*0.25*

Evaluation measures: Precision – Recall

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

Ranking	
red cross	*15.66*
iron curtain	*15.23*
decimal point	*14.01*
coupon book	*13.83*
book author	*11.05*
arithmetic operation	*10.52*
paper feeder	*10.17*
new book	*10.09*
round table	*7.03*
new wave	*6.59*
gas station	*6.04*
system type	*3.54*
central part	*1.54*
and others	*0.54*
program in	*0.35*
level is	*0.25*

Evaluation measures: Precision – Recall

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

Ranking	
red cross	15.66
iron curtain	15.23
decimal point	14.01
coupon book	13.83
book author	11.05
arithmetic operation	10.52
paper feeder	10.17
new book	10.09
round table	7.03
new wave	6.59
gas station	6.04
system type	3.54
central part	1.54
and others	0.54
program in	0.35
level is	0.25

Classification	
red cross	1
iron curtain	1
decimal point	1
coupon book	1
book author	0
arithmetic operation	0
paper feeder	0
new book	0
round table	0
new wave	0
gas station	0
system type	0
central part	0
and others	0
program in	0
level is	0

Evaluation measures: Precision – Recall

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

Ranking	
red cross	15.66
iron curtain	15.23
decimal point	14.01
coupon book	13.83
book author	11.05
arithmetic operation	10.52
paper feeder	10.17
new book	10.09
round table	7.03
new wave	6.59
gas station	6.04
system type	3.54
central part	1.54
and others	0.54
program in	0.35
level is	0.25

Classification	
red cross	1
iron curtain	1
decimal point	1
coupon book	1
book author	0
arithmetic operation	0
paper feeder	0
new book	0
round table	0
new wave	0
gas station	0
system type	0
central part	0
and others	0
program in	0
level is	0

Evaluation measures: Precision – Recall

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

Ranking	
red cross	15.66
iron curtain	15.23
decimal point	14.01
coupon book	13.83
book author	11.05
arithmetic operation	10.52
paper feeder	10.17
new book	10.09
round table	7.03
new wave	6.59
gas station	6.04
system type	3.54
central part	1.54
and others	0.54
program in	0.35
level is	0.25

Classification	
red cross	1
iron curtain	1
decimal point	1
coupon book	1
book author	0
arithmetic operation	0
paper feeder	0
new book	0
round table	0
new wave	0
gas station	0
system type	0
central part	0
and others	0
program in	0
level is	0

Precision	Recall
100 %	50 %

Evaluation measures: Precision – Recall

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

Ranking	
red cross	15.66
iron curtain	15.23
decimal point	14.01
coupon book	13.83
book author	11.05
arithmetic operation	10.52
paper feeder	10.17
new book	10.09
round table	7.03
new wave	6.59
gas station	6.04
system type	3.54
central part	1.54
and others	0.54
program in	0.35
level is	0.25

Classification	
red cross	1
iron curtain	1
decimal point	1
coupon book	1
book author	1
arithmetic operation	0
paper feeder	0
new book	0
round table	0
new wave	0
gas station	0
system type	0
central part	0
and others	0
program in	0
level is	0

Precision	Recall
100 %	50 %
80 %	50 %

Evaluation measures: Precision – Recall

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

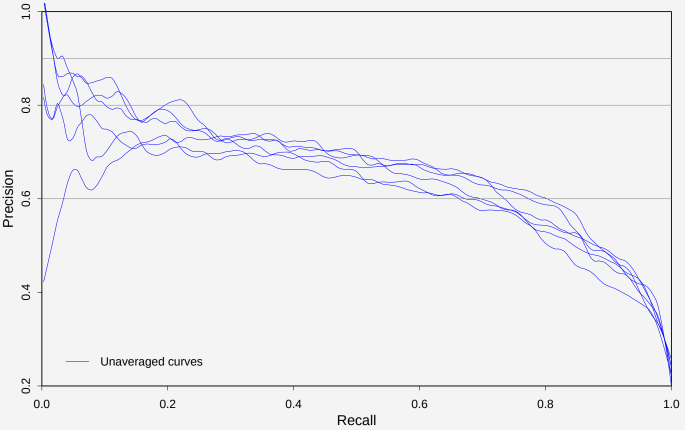
Ranking		Classification		Precision	Recall
red cross	15.66	red cross	1	100 %	12 %
iron curtain	15.23	iron curtain	1	100 %	25 %
decimal point	14.01	decimal point	1	100 %	37 %
coupon book	13.83	coupon book	1	100 %	50 %
book author	11.05	book author	1	80 %	50 %
arithmetic operation	10.52	arithmetic operation	1	83 %	62 %
paper feeder	10.17	paper feeder	1	85 %	75 %
new book	10.09	new book	1	75 %	75 %
round table	7.03	round table	1	77 %	87 %
new wave	6.59	new wave	1	70 %	87 %
gas station	6.04	gas station	1	72 %	100 %
system type	3.54	system type	1	66 %	100 %
central part	1.54	central part	1	61 %	100 %
and others	0.54	and others	1	57 %	100 %
program in	0.35	program in	1	53 %	100 %
level is	0.25	level is	1	50 %	100 %

- measured within the entire interval of possible threshold values

Visual evaluation: Precision-Recall curves

- graphical plots of recall vs. precision
- the closer to the top and right, the better ranking performance
- estimated for each eval fold and vertically averaging

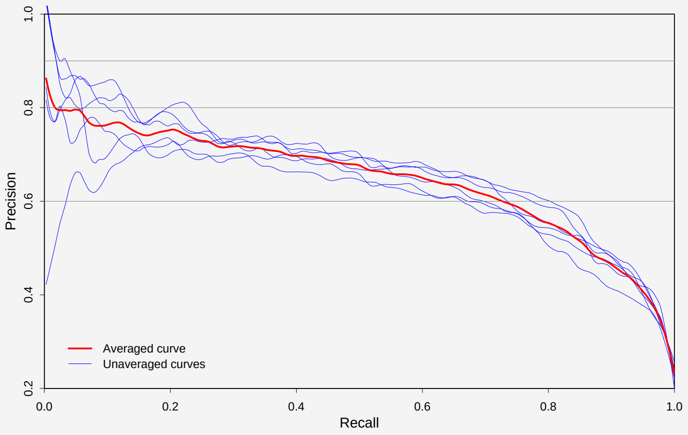
Precision-Recall curve averaging



Visual evaluation: Precision-Recall curves

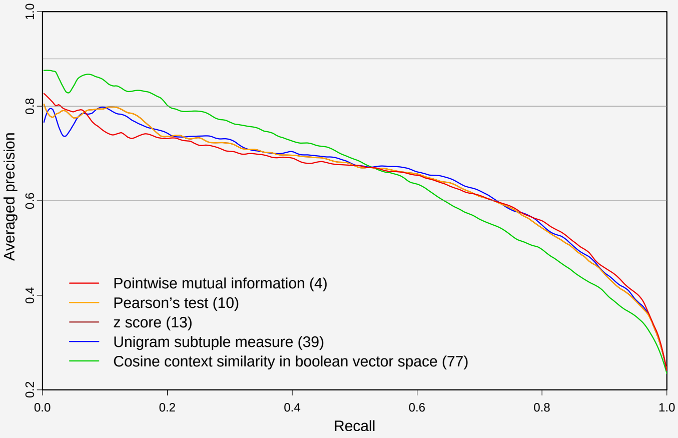
- graphical plots of recall vs. precision
- the closer to the top and right, the better ranking performance
- estimated for each eval fold and vertically averaging

Precision-Recall curve averaging



Evaluation results: Precision-Recall curves

The best-performing association measures



Evaluation measure: Average Precision

2) *Average Precision:* $E[P(R)],\, R \sim U(0,1)$ $\quad AP = \dfrac{1}{r}\sum_{i=1}^{r} p_i$

Ranking	
red cross	15.66
iron curtain	15.23
decimal point	14.01
coupon book	13.83
book author	11.05
arithmetic operation	10.52
paper feeder	10.17
new book	10.09
round table	7.03
new wave	6.59
gas station	6.04
system type	3.54
central part	1.54
and others	0.54
program in	0.35
level is	0.25

Classification	
red cross	1
iron curtain	1
decimal point	1
coupon book	1
book author	1
arithmetic operation	1
paper feeder	1
new book	1
round table	1
new wave	1
gas station	1
system type	1
central part	1
and others	1
program in	1
level is	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 measure: Average Precision

2) *Average Precision:* $E[P(R)],\ R \sim U(0,1)$ $$AP = \frac{1}{r}\sum_{i=1}^{r} p_i$$

Ranking		Classification		Precision	Recall
red cross	15.66	red cross	1	100 %	12 %
iron curtain	15.23	iron curtain	1	100 %	25 %
decimal point	14.01	decimal point	1	100 %	37 %
coupon book	13.83	coupon book	1	100 %	50 %
book author	11.05	book author	1	80 %	50 %
arithmetic operation	10.52	arithmetic operation	1	83 %	62 %
paper feeder	10.17	paper feeder	1	85 %	75 %
new book	10.09	new book	1	75 %	75 %
round table	7.03	round table	1	77 %	87 %
new wave	6.59	new wave	1	70 %	87 %
gas station	6.04	gas station	1	72 %	100 %
system type	3.54	system type	1	66 %	100 %
central part	1.54	central part	1	61 %	100 %
and others	0.54	and others	1	57 %	100 %
program in	0.35	program in	1	53 %	100 %
level is	0.25	level is	1	50 %	100 %

Evaluation measure: Average Precision

2) *Average Precision:* $E[P(R)], R \sim U(0,1)$ $AP = \frac{1}{r}\sum_{i=1}^{r} p_i$

Ranking		Classification		Precision	Recall
red cross	15.66	red cross	1	100 %	12 %
iron curtain	15.23	iron curtain	1	100 %	25 %
decimal point	14.01	decimal point	1	100 %	37 %
coupon book	13.83	coupon book	1	100 %	50 %
book author	11.05	book author	1	80 %	50 %
arithmetic operation	10.52	arithmetic operation	1	83 %	62 %
paper feeder	10.17	paper feeder	1	85 %	75 %
new book	10.09	new book	1	75 %	75 %
round table	7.03	round table	1	77 %	87 %
new wave	6.59	new wave	1	70 %	87 %
gas station	6.04	gas station	1	72 %	100 %
system type	3.54	system type	1	66 %	100 %
central part	1.54	central part	1	61 %	100 %
and others	0.54	and others	1	57 %	100 %
program in	0.35	program in	1	53 %	100 %
level is	0.25	level is	1	50 %	100 %

89.6 % = AP

Evaluation measure: Average Precision

2) *Average Precision:* $E[P(R)],\ R \sim U(0,1)$ $AP = \frac{1}{r}\sum_{i=1}^{r} p_i$

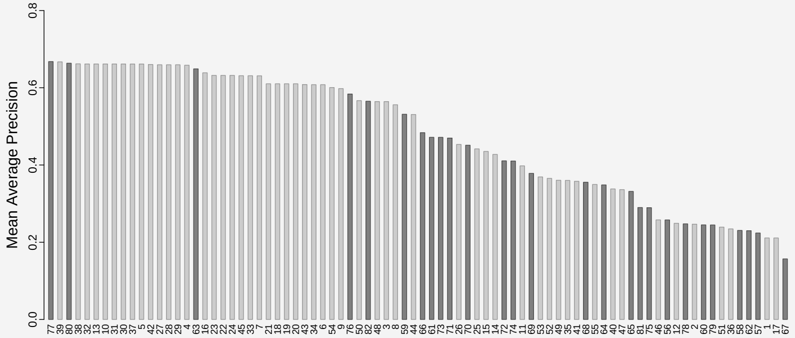
Ranking		Classification		Precision	Recall
red cross	*15.66*	*red cross*	*1*	*100 %*	*12 %*
iron curtain	*15.23*	*iron curtain*	*1*	*100 %*	*25 %*
decimal point	*14.01*	*decimal point*	*1*	*100 %*	*37 %*
coupon book	*13.83*	*coupon book*	*1*	*100 %*	*50 %*
book author	*11.05*	*book author*	*1*	*80 %*	*50 %*
arithmetic operation	*10.52*	*arithmetic operation*	*1*	*83 %*	*62 %*
paper feeder	*10.17*	*paper feeder*	*1*	*85 %*	*75 %*
new book	*10.09*	*new book*	*1*	*75 %*	*75 %*
round table	*7.03*	*round table*	*1*	*77 %*	*87 %*
new wave	*6.59*	*new wave*	*1*	*70 %*	*87 %*
gas station	*6.04*	*gas station*	*1*	*72 %*	*100 %*
system type	*3.54*	*system type*	*1*	*66 %*	*100 %*
central part	*1.54*	*central part*	*1*	*61 %*	*100 %*
and others	*0.54*	*and others*	*1*	*57 %*	*100 %*
program in	*0.35*	*program in*	*1*	*53 %*	*100 %*
level is	*0.25*	*level is*	*1*	*50 %*	*100 %*

3) *Mean Average Precision:* $E[AP]$ $MAP = \frac{1}{6}\sum_{i=1}^{6} AP_i$

89.6 % = AP

Overall results: Mean Average Precision

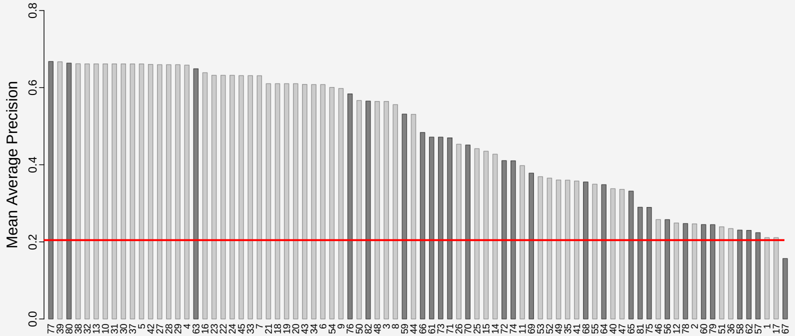
MAP of all lexical association measures in descending order



- Baseline (*ratio of true collocations*): 21.02 %
- Best context-based measure (■): Cosine similarity in vector space: 66.79 %
- Best statistical association measure (■): Unigram subtuple measure: 66.72 %
- Best 16 measures – statistically indistinguishable MAP ∼ current state of the art

Overall results: Mean Average Precision

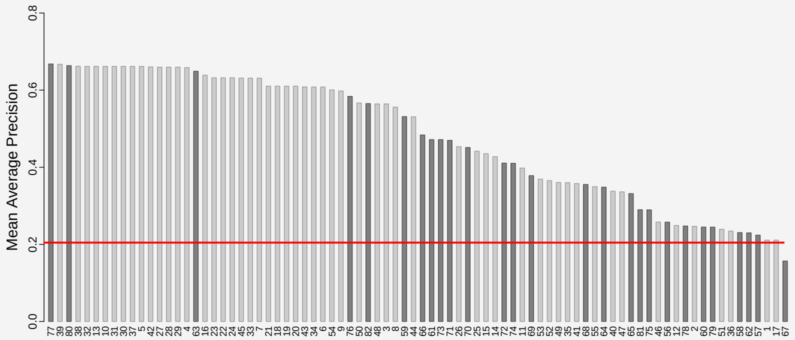
MAP of all lexical association measures in descending order



- Baseline (*ratio of true collocations*): 21.02 %
- Best context-based measure (■): Cosine similarity in vector space: 66.79 %
- Best statistical association measure (■): Unigram subtuple measure: 66.72 %
- Best 16 measures – statistically indistinguishable MAP ∼ current state of the art

Overall results: Mean Average Precision

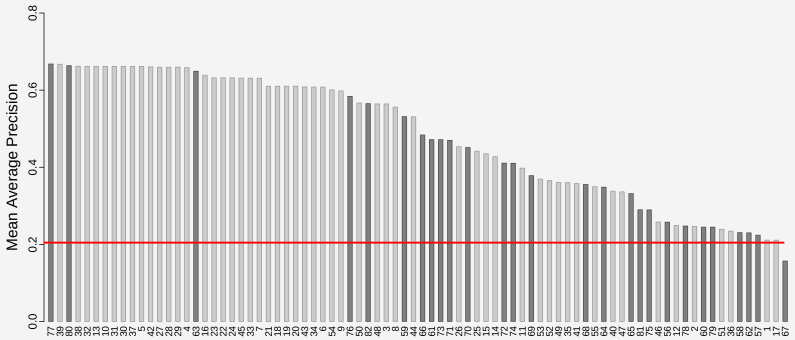
MAP of all lexical association measures in descending order



- Baseline (*ratio of true collocations*): 21.02 %
- Best context-based measure (■): Cosine similarity in vector space: 66.79 %
- Best statistical association measure (■): Unigram subtuple measure: 66.72 %
- Best 16 measures – statistically indistinguishable MAP ∼ current state of the art

Overall results: Mean Average Precision

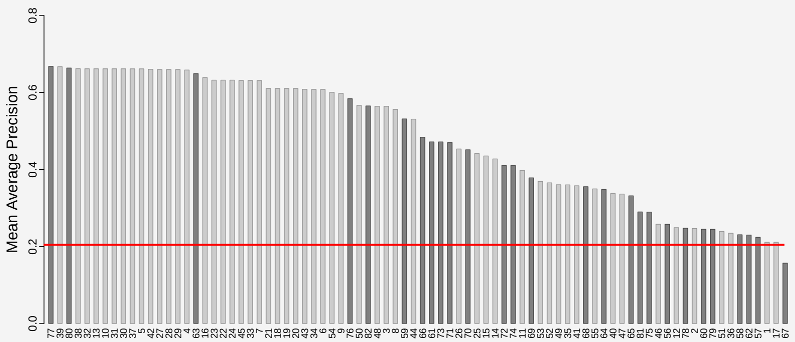
MAP of all lexical association measures in descending order



- Baseline (*ratio of true collocations*): 21.02 %
- Best context-based measure (■): Cosine similarity in vector space: 66.79 %
- Best statistical association measure (■): Unigram subtuple measure: 66.72 %
- Best 16 measures – statistically indistinguishable MAP ∼ current state of the art

Overall results: Mean Average Precision

MAP of all lexical association measures in descending order



- Baseline (*ratio of true collocations*): 21.02 %
- Best context-based measure (■): Cosine similarity in vector space: 66.79 %
- Best statistical association measure (■): Unigram subtuple measure: 66.72 %
- Best 16 measures – statistically indistinguishable MAP ∼ current state of the art

Combining association measures

Motivation

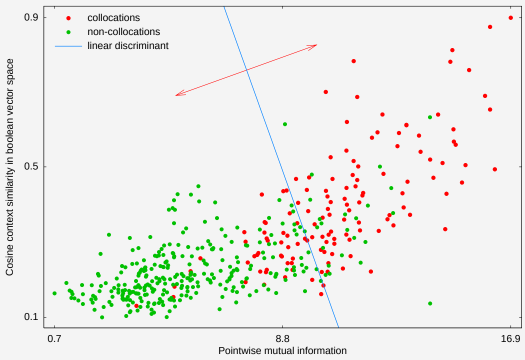
- different association measures discover different groups/types of collocations
- existence of uncorrelated association measures

Combining association measures

Motivation

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

5 % data sample from PDT-Dep

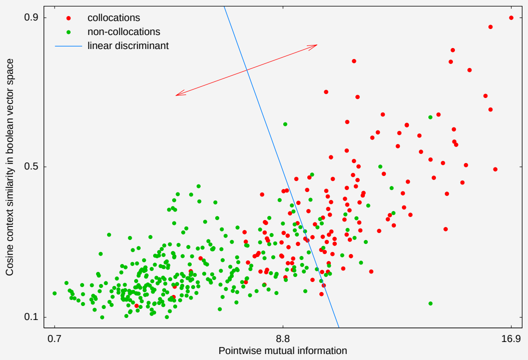


Combining association measures

Motivation

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

5 % data sample from PDT-Dep



Note: So far all methods – unsupervised, the combination methods – supervised

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

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

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

Combination models: Evaluation

Evaluation scheme

- 6-fold crossvalidation on the 6 evaluation folds
- 5 folds for training (*fitting parameters*), 1 fold for testing (*ranking performance*)
- PR curve and AP score estimated on each test fold and averaged

train$_1$	***train$_2$***	***train$_3$***	***train$_4$***	***train$_5$***	***test$_6$***	***held-out***

Results: Mean Average Precision

method	*MAP*	*+%*
Unigram subtuple measure	66.72	–
Cosine similarity in vector space	66.79	0.00
Support Vector Machine	73.03	9.35
Neural Network (1 unit)	74.88	12.11
Linear Discriminant Analysis	75.16	12.54
Linear Logistic Regression	77.36	15.82
Neural Network (5 units)	80.87	21.08

Combination models: Evaluation

Evaluation scheme

- 6-fold crossvalidation on the 6 evaluation folds
- 5 folds for training (*fitting parameters*), 1 fold for testing (*ranking performance*)
- PR curve and AP score estimated on each test fold and averaged

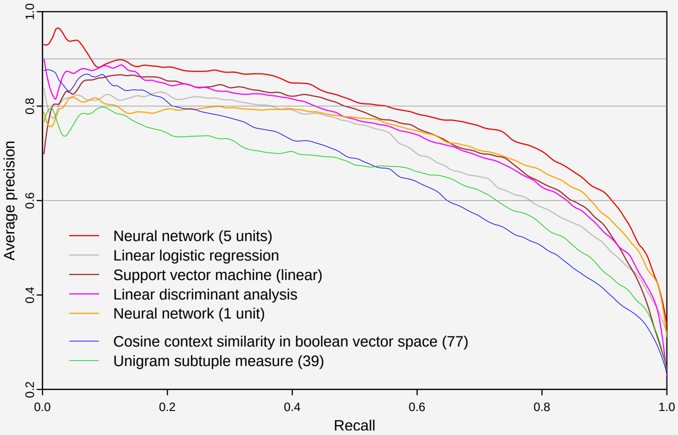
$train_1$	$train_2$	$train_3$	$train_4$	$train_5$	$test_6$	held-out

Results: Mean Average Precision

method	*MAP*	*+%*
Unigram subtuple measure	66.72	–
Cosine similarity in vector space	66.79	0.00
Support Vector Machine	73.03	9.35
Neural Network (1 unit)	74.88	12.11
Linear Discriminant Analysis	75.16	12.54
Linear Logistic Regression	77.36	15.82
Neural Network (5 units)	80.87	21.08

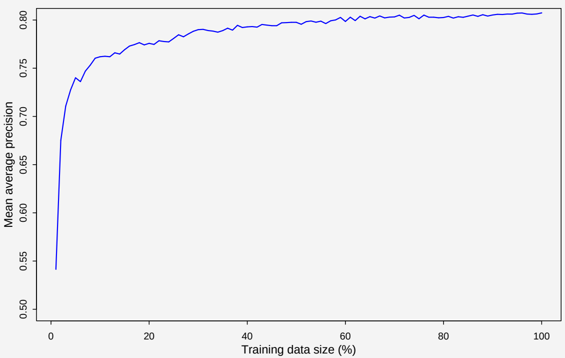
Results: Precision-Recall curves

Combination methods compared with the best association measures



Learning curve analysis

Neural network (5 units) learning curve



- 100% of training data = 5 training folds (8 737 annotated collocation candidates)
- 95% of the final MAP achieved with 15% of training data
- 99% of the final MAP achieved with 50% of training data

Adding linguistic features

Idea

- improving the combination models by adding linguistic features
- categorical features can be transformed to binary dummy features

New features

- Part-of-Speech pattern: combination of component POS (*A:N, N:N, . . .*)
- Syntactic relation: dependency type (*attribute, object, . . .*)

Results: Mean Average Precision

method	*MAP*	*+%*
Unigram subtuple measure	66.72	–
Cosine similarity in vector space	66.79	0.00
NNet/5 (AM)	80.87	21.08
NNet/5 (AM+POS)	82.79	24.09
NNet/5 (AM+POS+DEP)	84.53	26.69

Adding linguistic features

Idea

- improving the combination models by adding linguistic features
- categorical features can be transformed to binary dummy features

New features

- Part-of-Speech pattern: combination of component POS (*A:N, N:N, . . .*)
- Syntactic relation: dependency type (*attribute, object, . . .*)

Results: Mean Average Precision

method	*MAP*	*+%*
Unigram subtuple measure	66.72	–
Cosine similarity in vector space	66.79	0.00
NNet/5 (AM)	80.87	21.08
NNet/5 (AM+POS)	82.79	24.09
NNet/5 (AM+POS+DEP)	84.53	26.69

Adding linguistic features

Idea

- improving the combination models by adding linguistic features
- categorical features can be transformed to binary dummy features

New features

- Part-of-Speech pattern: combination of component POS (*A:N, N:N, . . .*)
- Syntactic relation: dependency type (*attribute, object, . . .*)

Results: Mean Average Precision

method	*MAP*	*+%*
Unigram subtuple measure	66.72	–
Cosine similarity in vector space	66.79	0.00
NNet/5 (AM)	80.87	21.08
NNet/5 (AM+POS)	82.79	24.09
NNet/5 (AM+POS+DEP)	84.53	26.69

Model reduction

Motivation

- *"Ocama's razor"*
- combination of all 82 association measures is too complex
- models should be reduced: redundant variables removed

Two issues

1. groups of highly correlated measures
2. measures with no or minimal contribution to the model

Two-step solution

1. correlation based clustering; one representative selected from each cluster
2. step-wise procedure removing variables one by one

Model reduction

Motivation

- *"Ocama's razor"*
- combination of all 82 association measures is too complex
- models should be reduced: redundant variables removed

Two issues

1. groups of highly correlated measures
2. measures with no or minimal contribution to the model

Two-step solution

1. correlation based clustering; one representative selected from each cluster
2. step-wise procedure removing variables one by one

Model reduction

Motivation

- *"Ocama's razor"*
- combination of all 82 association measures is too complex
- models should be reduced: redundant variables removed

Two issues

1. groups of highly correlated measures
2. measures with no or minimal contribution to the model

Two-step solution

1. correlation based clustering; one representative selected from each cluster
2. step-wise procedure removing variables one by one

Model reduction: 1) Clustering

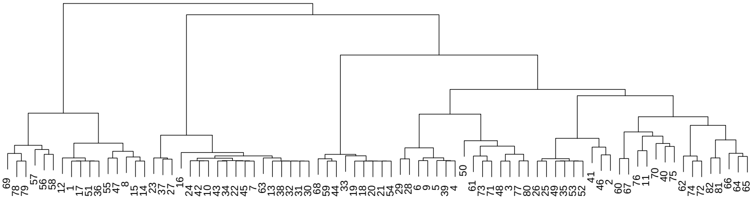
Agglomerative hierarchical clustering

- groups the measures with the same/similar contribution to the model
- begins with each measure as a separate cluster and merge them into successively larger clusters
- distance metrics = 1- |*Pearson's correlation*| (estimated on the *held-out* fold)

Model reduction: 1) Clustering

Agglomerative hierarchical clustering

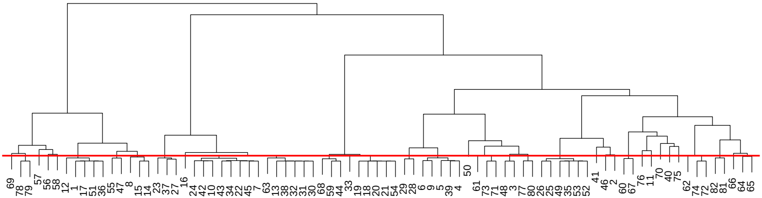
- groups the measures with the same/similar contribution to the model
- begins with each measure as a separate cluster and merge them into successively larger clusters
- distance metrics = 1- |*Pearson's correlation*| (estimated on the *held-out* fold)



Model reduction: 1) Clustering

Agglomerative hierarchical clustering

- groups the measures with the same/similar contribution to the model
- begins with each measure as a separate cluster and merge them into successively larger clusters
- distance metrics = 1- |*Pearson's correlation*| (estimated on the *held-out* fold)



- number of the final clusters empirically set to 60
- the best performing measure (by MAP on the *held-out* fold) selected as the representative from each cluster

Model reduction: 2) Stepwise variable removal

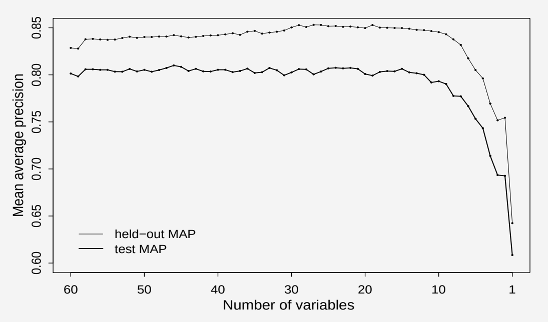
Iterative procedure

- initiated with the 60 variables/measures
- in each iteration we remove the variable causing minimal performance degradation when not used in the model (by MAP on the *held-out* fold)
- stops before the degradation becomes statistically significant

Model reduction: 2) Stepwise variable removal

Iterative procedure

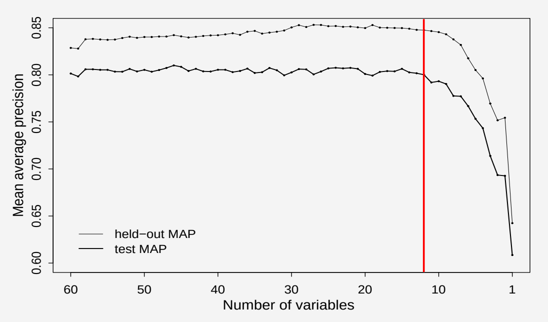
- initiated with the 60 variables/measures
- in each iteration we remove the variable causing minimal performance degradation when not used in the model (by MAP on the *held-out* fold)
- stops before the degradation becomes statistically significant



Model reduction: 2) Stepwise variable removal

Iterative procedure

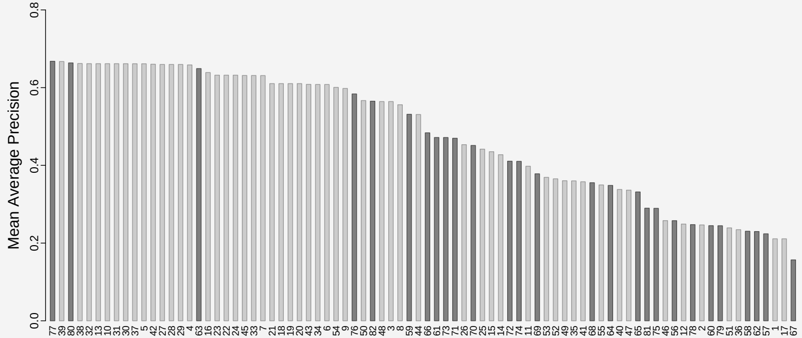
- initiated with the 60 variables/measures
- in each iteration we remove the variable causing minimal performance degradation when not used in the model (by MAP on the *held-out* fold)
- stops before the degradation becomes statistically significant



- the final model contains 13 variables/lexical association measures

Model reduction: Process overview

MAP of individual lexical association measures

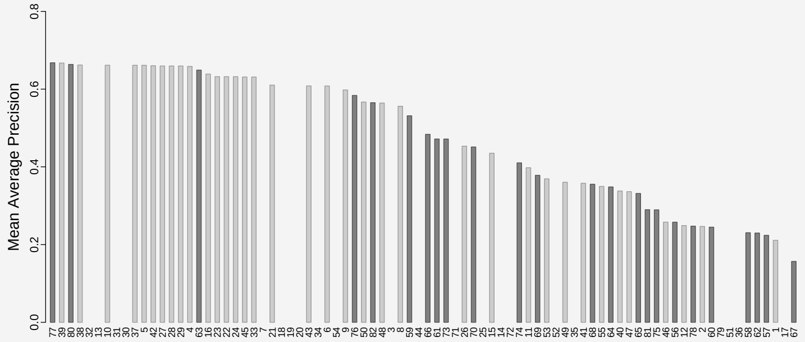


- procedure initiated with all 82 association measures
- highly correlated measures removed in the first phase (*clustering*)
- 13 measures left after the second phase (*stepwise removal*)

  = 4 statistical association measures (■) + 9 context-based measures (■)

Model reduction: Process overview

MAP of individual lexical association measures

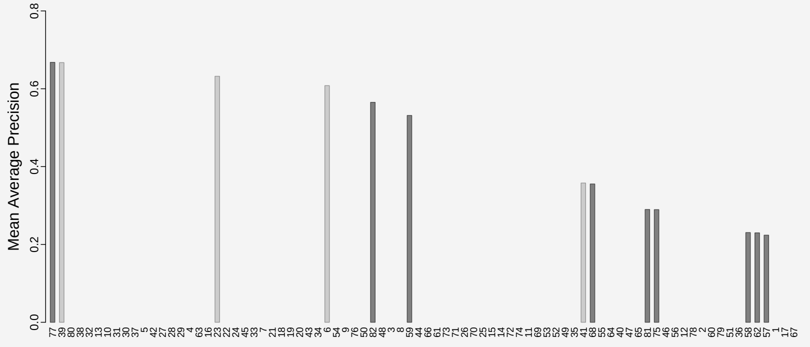


- procedure initiated with all 82 association measures
- highly correlated measures removed in the first phase (*clustering*)
- 13 measures left after the second phase (*stepwise removal*)

  = 4 statistical association measures (■) + 9 context-based measures (■)

Model reduction: Process overview

MAP of individual lexical association measures

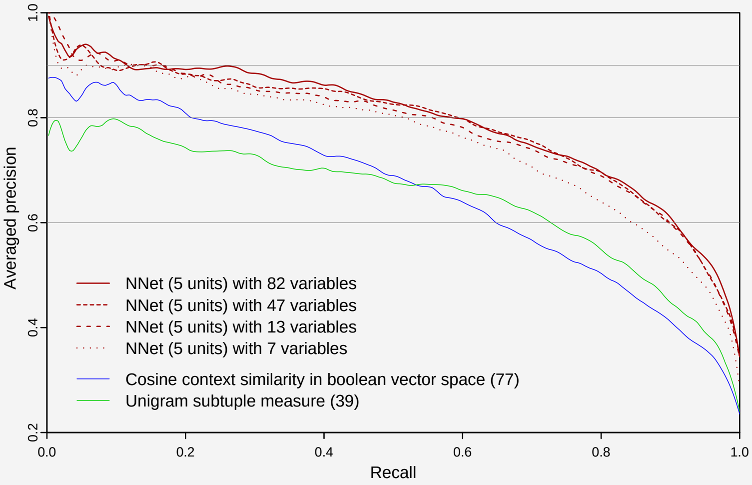


- procedure initiated with all 82 association measures
- highly correlated measures removed in the first phase (*clustering*)
- 13 measures left after the second phase (*stepwise removal*)

  = 4 statistical association measures (■) + 9 context-based measures (■)

Model reduction results: Precision-Recall curves

Reduced combination models compared with the best association measures



Conslusions

Main results

1. inventory of 82 lexical association measures
2. 4 reference data sets
3. all lexical association measures evaluated on these data sets
4. combining association measures improved *state of the art* in collocation extraction
5. combination models reduced to 13 measures without performance degradation

Other contribution of the thesis

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

Conslusions

Main results

1. inventory of 82 lexical association measures
2. 4 reference data sets
3. all lexical association measures evaluated on these data sets
4. combining association measures improved *state of the art* in collocation extraction
5. combination models reduced to 13 measures without performance degradation

Other contribution of the thesis

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

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*, 44:137-158, 2010.

- Pavel Pecina: **Lexical Association Measures: Collocation Extraction**, *volume 4 of Studies in Computational and Theoretical Linguistics*, UFAL, Praha, Czech Republic, 2009.

- 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.

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
- annotation consistent with *PDT-Surf*

PAR-Dist

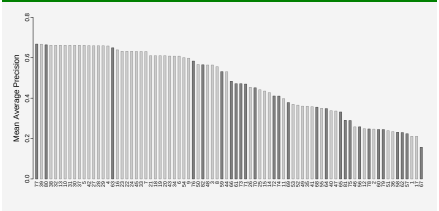
- source corpus: Swedish Parole, 22 mil. tokens
- automatic morphological tagging and lemmatization
- distance bigrams: word pairs occurring within a distance of 1–3 words
- annotation: non-exhaustive manual extraction of support verb constructions
- no frequency filter applied

Reference data summary

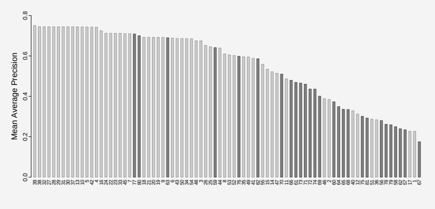
reference data set	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*	*PAR-Dist*
source corpus	*PDT*	*PDT*	*CNC*	*PAROLE*
language	*Czech*	*Czech*	*Czech*	*Swedish*
morphology	*manual*	*manual*	*auto*	*auto*
syntax	*manual*	*none*	*none*	*none*
bigram types	*dependency*	*surface*	*surface*	*distance*
tokens	1 504 847	1 504 847	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

Context-based vs. statistical association measures

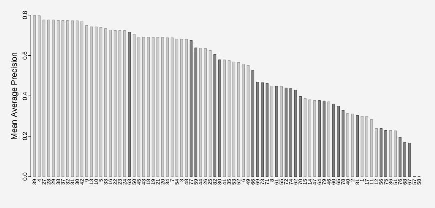
PDT-Dep



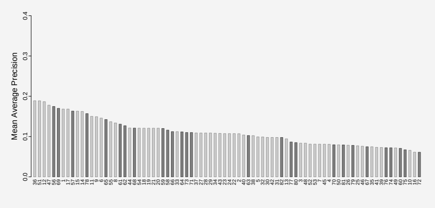
PDT-Surf



CNC-Surf

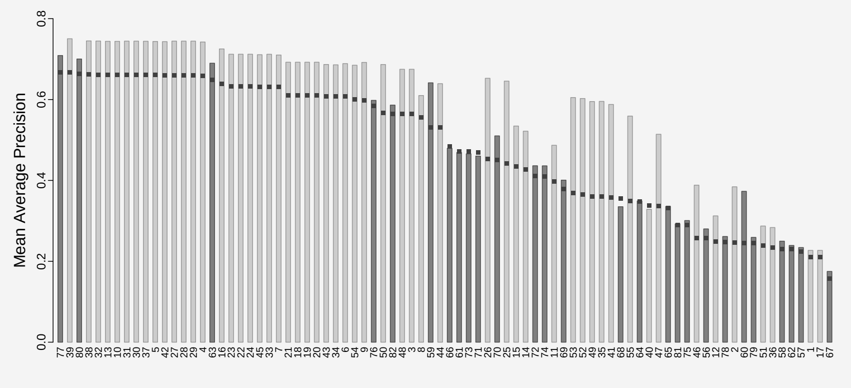


PAR-Dist



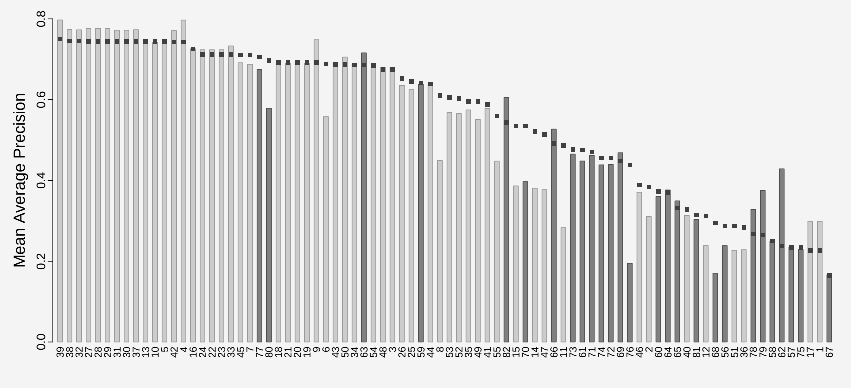
Results / Mean average precision: *PDT-Dep* vs. *PDT-Surf*

Dependency bigrams vs. surface bigrams



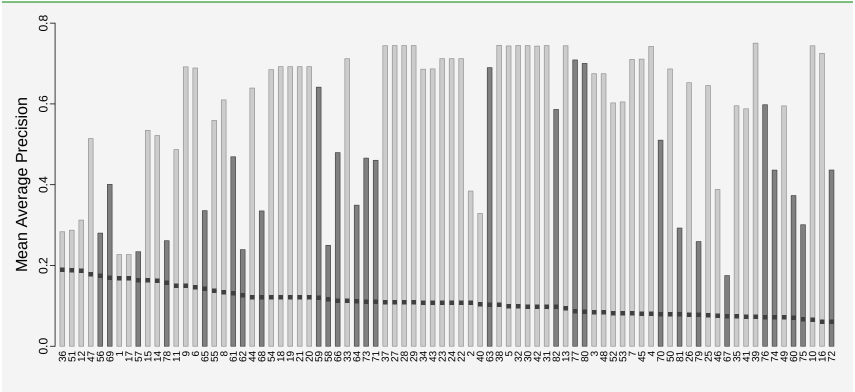
Results / Mean average precision: *PDT-Surf* vs. *CNC-Surf*

Small source corpus vs. large source corpus



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

Different corpus, different language, different task



Comparison of AM evaluation results on different data sets