

# Lexical Association Measures

## Collocation Extraction

Pavel Pecina

`pecina@ufal.mff.cuni.cz`

Institute of Formal and Applied Linguistics  
Charles University, Prague



ÚČNK  
7.2.2012

## Talk outline

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

# Lexical association

1/30

## 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<sub>(FR)</sub> – house<sub>(EN)</sub>, baum<sub>(GE)</sub> – tree<sub>(EN)</sub>, květina<sub>(CZ)</sub> – flower<sub>(EN)</sub>*

## 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

1/30

## 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<sub>(FR)</sub> – house<sub>(EN)</sub>, baum<sub>(GE)</sub> – tree<sub>(EN)</sub>, květina<sub>(CZ)</sub> – flower<sub>(EN)</sub>*

## 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

1/30

## 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<sub>(FR)</sub> – house<sub>(EN)</sub>, baum<sub>(GE)</sub> – tree<sub>(EN)</sub>, květina<sub>(CZ)</sub> – flower<sub>(EN)</sub>*

## 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

2/30

## 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

2/30

## 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

2/30

## 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

3/30

## 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

3/30

### 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

3/30

### 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

4/30

### 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
- ▶ ranges from **free word combinations** to **idioms**
- ▶ specified **intensionally** (general rules) or **extensionally** (particular constraints)

### 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

1. idioms (*to kick the bucket, to hear st. through the grapevine*)
2. proper names (*New York, Old Town, Vaclav Havel*)
3. technical terms (*car oil, stock owl, hard disk*)
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*)

# Collocational association

4/30

## 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
- ▶ ranges from **free word combinations** to **idioms**
- ▶ specified **intensionally** (general rules) or **extensionally** (particular constraints)

## 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

1. idioms (*to kick the bucket, to hear st. through the grapevine*)
2. proper names (*New York, Old Town, Vaclav Havel*)
3. technical terms (*car oil, stock owl, hard disk*)
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*)

# Collocational association

4/30

## 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
- ▶ ranges from **free word combinations** to **idioms**
- ▶ specified **intensionally** (general rules) or **extensionally** (particular constraints)

## 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

1. idioms (*to kick the bucket, to hear st. through the grapevine*)
2. proper names (*New York, Old Town, Vaclav Havel*)
3. technical terms (*car oil, stock owl, hard disk*)
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 properties

5/30

### 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 \*big church mouse*

### Lexical non-substitutability

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

### Translatability into other languages

- ▶ translation cannot generally be performed blindly, word by word  
*ice cream – zmrzlina*

---

### Domain dependency

- ▶ collocational character only in specific domains  
*carriage return*

## Collocation properties

5/30

### 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 \*big church mouse*

### Lexical non-substitutability

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

### Translatability into other languages

- ▶ translation cannot generally be performed blindly, word by word  
*ice cream – zmrzlina*

---

### Domain dependency

- ▶ collocational character only in specific domains  
*carriage return*



## Collocation properties

5/30

### 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 \*big church mouse*

### Lexical non-substitutability

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

### Translatability into other languages

- ▶ translation cannot generally be performed blindly, word by word  
*ice cream – zmrzlina*

---

### Domain dependency

- ▶ collocational character only in specific domains  
*carriage return*

## Collocation properties

5/30

### 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 \*big church mouse*

### Lexical non-substitutability

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

### Translatability into other languages

- ▶ translation cannot generally be performed blindly, word by word  
*ice cream – zmrzlina*

---

### Domain dependency

- ▶ collocational character only in specific domains  
*carriage return*

## Collocation properties

5/30

### 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 \*big church mouse*

### Lexical non-substitutability

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

### Translatability into other languages

- ▶ translation cannot generally be performed blindly, word by word  
*ice cream – zmrzlina*
- 

### Domain dependency

- ▶ collocational character only in specific domains  
*carriage return*

## Collocation extraction

6/30

### 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 occurrences 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

6/30

## 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 occurrences 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

6/30

### 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 occurrences 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

7/30

“Collocation components occur together more often than by chance”

- ▶ the corpus is interpreted 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,  $\chi^2$  test, Odds ratio, Jaccard, Pointwise mutual information*

Example: *Pointwise Mutual Information*

Data:  $f(\text{iron curtain}) = 11$

$f(\text{iron}) = 30$

$f(\text{curtain}) = 15$

MLE:  $p(\text{iron curtain}) = 0.000007$

$p(\text{iron}) = 0.000020$

$p(\text{curtain}) = 0.000010$

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

$\hat{f}(\text{iron curtain}) = 0.000030$

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

## Extraction principle I

7/30

“Collocation components occur together more often than by chance”

- ▶ the corpus is interpreted 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,  $\chi^2$  test, Odds ratio, Jaccard, Pointwise mutual information*

## Example: Pointwise Mutual Information

Data:  $f(\text{iron curtain}) = 11$

$f(\text{iron}) = 30$

$f(\text{curtain}) = 15$

MLE:  $p(\text{iron curtain}) = 0.000007$

$p(\text{iron}) = 0.000020$

$p(\text{curtain}) = 0.000010$

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

$\hat{f}(\text{iron curtain}) = 0.000030$

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



## Extraction principle II

8/30

## “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:  $\mathbf{p}(w|C_{xy}^r), \mathbf{p}(w|C_{xy}^l)$
  2. uncertainty (**entropy**) what the next/previous word is:  $H(\mathbf{p}(w|C_{xy}^r)), H(\mathbf{p}(w|C_{xy}^l))$
- ▶ 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:  $H(\mathbf{p}(w|C_{xy}^r))$



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

## Extraction principle II

8/30

## “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:  $\mathbf{p}(w|C_{xy}^r), \mathbf{p}(w|C_{xy}^l)$
  2. uncertainty (**entropy**) what the next/previous word is:  $H(\mathbf{p}(w|C_{xy}^r)), H(\mathbf{p}(w|C_{xy}^l))$
- ▶ 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:  $H(\mathbf{p}(w|C_{xy}^r))$



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 krizi vstoupil do Bělehradu **černý trh**, pašování a zvýšená kriminalita. Pfkupníci provázejí ...  
 ... 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, miní. 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. **Maltés** 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 provázejí ...  
 ... 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, miní. 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	$P(x,y)$
2.	Conditional probability	$P(x y)$
3.	Reverse conditional probability	$P(y x)$
4.	Pointwise mutual information	$\log \frac{P(x,y)}{P(x)P(y)}$
5.	Mutual dependency (MD)	$\log \frac{P(x,y)}{P(x)P(y)}$
6.	Log frequency biased MD	$\log \frac{P(x,y)}{P(x)P(y)} + \log P(x,y)$
7.	Normalized expectation	$\frac{f(x,y) - f(x)f(y)}{f(x) + f(y)}$
8.	Mutual expectation	$\frac{f(x,y) - f(x)f(y)}{f(x) + f(y)} + P(x,y)$
9.	Salience	$\log \frac{P(x,y)}{P(x)P(y)} - \log P(x,y)$
10.	Pearson's $\chi^2$ test	$\sum_x \frac{(f_x - f_x^e)^2}{f_x}$
11.	Fisher's exact test	$\frac{f(x,y)(f(x)-f(x,y))(f(y)-f(x,y))}{f(x)f(y)}$
12.	t test	$\frac{f(x,y) - (f(x)f(y)/N)}{f(x) - f(x,y)}$
13.	z score	$\frac{f(x,y) - (f(x)f(y)/N)}{\sqrt{f(x)(1-f(x)/N)}}$
14.	Poisson significance measure	$\frac{f(x,y) - f(x)f(y)}{f(x) - f(x)f(y)}$
15.	Log likelihood ratio	$-2 \sum_{i,j} f_{ij} \log f_{ij} / f_{i.} f_{.j}$
16.	Squared log likelihood ratio	$-2 \sum_{i,j} \log f_{ij}^2 / f_{i.} f_{.j}$
17.	Russel-Rao	$\frac{f(x,y)}{f(x) + f(y) - f(x,y)}$
18.	Sokal-Michiner	$\frac{f(x,y)}{f(x) + f(y)}$
19.	Rogers-Tanimoto	$\frac{f(x,y)}{f(x) + f(y) - f(x,y)}$
20.	Hamann	$\frac{f(x,y)}{f(x) + f(y)}$
21.	Third Sokal-Sneath	$\frac{f(x,y)}{f(x) + f(y) - f(x,y)}$
22.	Jaccard	$\frac{f(x,y)}{f(x) + f(y) - f(x,y)}$
23.	First Kulczynski	$\frac{f(x,y)}{f(x) + f(y)}$
24.	Second Sokal-Sneath	$\frac{f(x,y)}{f(x) + f(y) - f(x,y)}$
25.	Second Kulczynski	$\frac{f(x,y)}{f(x) + f(y)}$
26.	Fourth Sokal-Sneath	$\frac{f(x,y)}{f(x) + f(y) - f(x,y)}$
27.	Odds ratio	$\frac{f(x,y) - f(x)f(y)}{f(x) - f(x,y)}$
28.	Yulle's $\omega$	$\frac{f(x,y) - f(x)f(y)}{f(x) + f(y)}$
29.	Yulle's Q	$\frac{f(x,y) - f(x)f(y)}{f(x) + f(y)}$
30.	Driver-Kroeber	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
31.	Fifth Sokal-Sneath	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
32.	Pearson	$\frac{f(x,y) - f(x)f(y)}{\sqrt{(f(x)+f(y))}}$
33.	Baroni-Urbani	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
34.	Braun-Blanquet	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
35.	Simpson	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
36.	Michael	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
37.	Moutiford	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
38.	Fager	$\frac{f(x,y)}{\sqrt{(f(x)+f(y))}}$
39.	Unigram subtuples	$\log \frac{f(x)}{f(x) - 3.29\sqrt{\frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}}}$
40.	U cost	$\log(1 + \frac{f(x,y)}{\max(f(x), f(y))})$
41.	S cost	$\log(1 + \frac{f(x,y)}{\max(f(x), f(y))})$
42.	R cost	$\log(1 + \frac{f(x,y)}{\max(f(x), f(y))})$
43.	T combined cost	$\sqrt{U \times S \times R}$
44.	Phi	$\frac{P(x,y) - P(x)P(y)}{\sqrt{(P(x)P(y)(1-P(x))(1-P(y)))}}$
45.	Kappa	$\frac{P(x,y) - P(x)P(y)}{P(x,y) + P(x)P(y) - P(x)P(y)}$

#	Name	Formula
46.	J measure	$\max\{P(x,y) \log \frac{P(x,y)}{P(x)P(y)} + P(x,y) \log \frac{P(x,y)}{P(x,y)}\}$
47.	Gini index	$\max\{P(x,y) \log \frac{P(x,y)}{P(x)P(y)} + P(x,y) \log \frac{P(x,y)}{P(x,y)}\}$
48.	Confidence	$\max\{P(x,y), P(x y)\}$
49.	Laplace	$\max\{\frac{P(x,y)}{P(x)}, \frac{P(x,y)}{P(y)}\}$
50.	Conviction	$\max\{\frac{P(x,y)}{P(x)}, \frac{P(x,y)}{P(y)}\}$
51.	Platersty-Shapiro	$P(x,y) - P(x)P(y)$
52.	Certainty factor	$\max\{\frac{P(x,y) - P(x)P(y)}{P(x)}, \frac{P(x,y) - P(x)P(y)}{P(y)}\}$
53.	Added value (AV)	$\max\{P(y) - P(x,y), P(x) - P(x,y)\}$
54.	Collective strength	$\frac{P(x,y) + P(x,y)}{P(x) + P(y) - P(x,y)}$
55.	Klogsen	$\sqrt{P(x,y) \cdot AV}$
56.	Context entropy	$-\sum_w P(w C_x) \log P(w C_x)$
57.	Left context entropy	$-\sum_w P(w C_x) \log P(w C_x)$
58.	Right context entropy	$-\sum_w P(w C_x) \log P(w C_x)$
59.	Left context divergence	$P(x) \log P(x) - \sum_w P(w C_x) \log P(w C_x)$
60.	Right context divergence	$P(y) \log P(y) - \sum_w P(w C_x) \log P(w C_x)$
61.	Cross entropy	$-\sum_w P(w C_x) \log P(w C_x)$
62.	Reverse cross entropy	$-\sum_w P(w C_x) \log P(w C_x)$
63.	Intersection measure	$\frac{P(x,y)}{P(x) + P(y)}$
64.	Euclidean norm	$\sqrt{\sum_w (P(w C_x) - P(w C_x))^2}$
65.	Cosine norm	$\frac{P(x,y)}{\sqrt{P(x)P(y)}}$
66.	L1 norm	$ \sum_w (P(w C_x) - P(w C_x)) $
67.	Confusion probability	$\sum_w \frac{P(w C_x) - P(w C_x)}{P(x)}$
68.	Reverse confusion probability	$\sum_w \frac{P(w C_x) - P(w C_x)}{P(y)}$
69.	Jensen-Shannon divergence	$\frac{1}{2} [D(p(w C_x)    \frac{1}{2}(p(w C_x) + p(w C_x)))] + D(p(w C_x)    \frac{1}{2}(p(w C_x) + p(w C_x)))]$
70.	Cosine of pointwise MI	$\frac{\sqrt{\sum_w MI(w)^2} \cdot \sqrt{\sum_w MI(w)^2}}{\sum_w P(w C_x) \log \frac{P(w C_x)}{P(w C_x)}}$
71.	KL divergence	$\sum_w P(w C_x) \log \frac{P(w C_x)}{P(w C_x)}$
72.	Reverse KL divergence	$\sum_w P(w C_x) \log \frac{P(w C_x)}{P(w C_x)}$
73.	Skew divergence	$D(p(w C_x)    \alpha p(w C_x) + (1-\alpha)p(w C_x))$
74.	Reverse skew divergence	$D(p(w C_x)    \alpha p(w C_x) + (1-\alpha)p(w C_x))$
75.	Phrase word cooccurrence	$\frac{1}{2} (\frac{f(x,y)}{f(x)} + \frac{f(x,y)}{f(y)})$
76.	Word association	$\frac{1}{2} (\frac{f(x,y)}{f(x)} + \frac{f(x,y)}{f(y)})$
Cosine context similarity:		$\frac{1}{2} (\cos(c_x, c_x) + \cos(c_x, c_x))$
		$c_x = (z_1); \cos(c_x, c_x) = \frac{1}{\sqrt{z_1^2 + z_1^2}}$
77.	in boolean vector space	$z_1 = \delta(f(w C_x))$
78.	in tf vector space	$z_1 = f(w C_x)$
79.	in tf · idf vector space	$z_1 = f(w C_x) \cdot \frac{1}{\sqrt{f(w C_x)}}$
Dice context similarity:		$\frac{1}{2} (\text{dice}(c_x, c_x) + \text{dice}(c_x, c_x))$
		$c_x = (z_1); \text{dice}(c_x, c_x) = \frac{1}{\sum_{i=1}^n z_i + \sum_{i=1}^n z_i}$
80.	in boolean vector space	$z_1 = \delta(f(w C_x))$
81.	in tf vector space	$z_1 = f(w C_x)$
82.	in tf · idf vector space	$z_1 = f(w C_x) \cdot \frac{1}{\sqrt{f(w C_x)}}$

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

## Extraction pipeline

11/30

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 chosen lexical association measure
6. **ranking/classification** of collocation candidates according to their scores

### Ranking

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

### Classification

<i>red cross</i>	1
<i>decimal point</i>	1
<i>arithmetic operation</i>	1
<i>paper feeder</i>	1
<i>system type</i>	0
<i>and others</i>	0
<i>program in</i>	0
<i>level is</i>	0

## Extraction pipeline

11/30

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 chosen lexical association measure
6. **ranking/classification** of collocation candidates according to their scores

## Ranking

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

## Classification

<i>red cross</i>	1
<i>decimal point</i>	1
<i>arithmetic operation</i>	1
<i>paper feeder</i>	1
<i>system type</i>	0
<i>and others</i>	0
<i>program in</i>	0
<i>level is</i>	0

## Extraction pipeline

11/30

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 chosen lexical association measure
6. **ranking/classification** of collocation candidates according to their scores

### Ranking

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

### Classification

<i>red cross</i>	1
<i>decimal point</i>	1
<i>arithmetic operation</i>	1
<i>paper feeder</i>	1
<hr/>	
<i>system type</i>	0
<i>and others</i>	0
<i>program in</i>	0
<i>level is</i>	0



## Reference data set

12/30

### 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*)
- ▶ 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

12/30

### 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*)
- ▶ 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 on *morphological* and *analytical* level

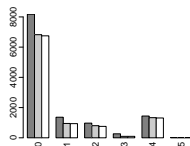
### Collocation candidates

- ▶ **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*)
- ▶ 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

13/30

### 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



### 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

13/30

### 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



### 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

13/30

### 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



### 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

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

## Evaluation measures: Precision – Recall

14/30

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

## Ranking

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



## Evaluation measures: Precision – Recall

14/30

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

## Ranking

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

## Evaluation measures: Precision – Recall

14/30

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

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

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

## Evaluation measures: Precision – Recall

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

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

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

## Evaluation measures: Precision – Recall

14/30

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

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

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

Precision	Recall
100 %	50 %

## Evaluation measures: Precision – Recall

14/30

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

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

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

Precision	Recall
100 %	50 %
80 %	50 %

## Evaluation measures: Precision – Recall

14/30

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

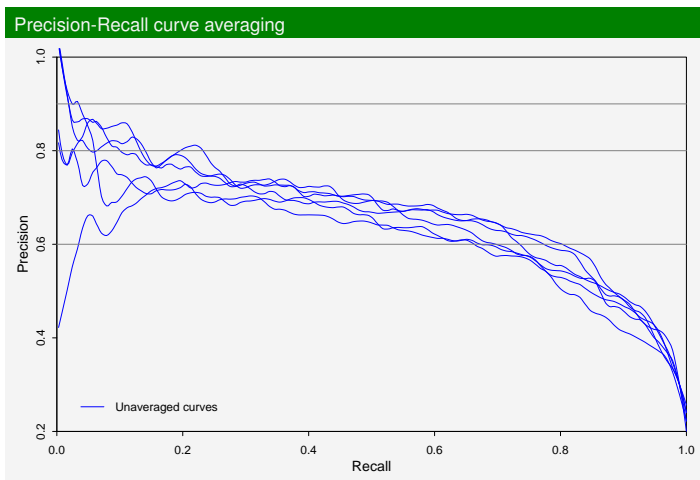
Ranking		Classification		Precision	Recall
<i>red cross</i>	15.66	<i>red cross</i>	1	100 %	12 %
<i>iron curtain</i>	15.23	<i>iron curtain</i>	1	100 %	25 %
<i>decimal point</i>	14.01	<i>decimal point</i>	1	100 %	37 %
<i>coupon book</i>	13.83	<i>coupon book</i>	1	100 %	50 %
<i>book author</i>	11.05	<i>book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

- ▶ measured within the entire interval of possible threshold values

## Visual evaluation: Precision-Recall curves

15/30

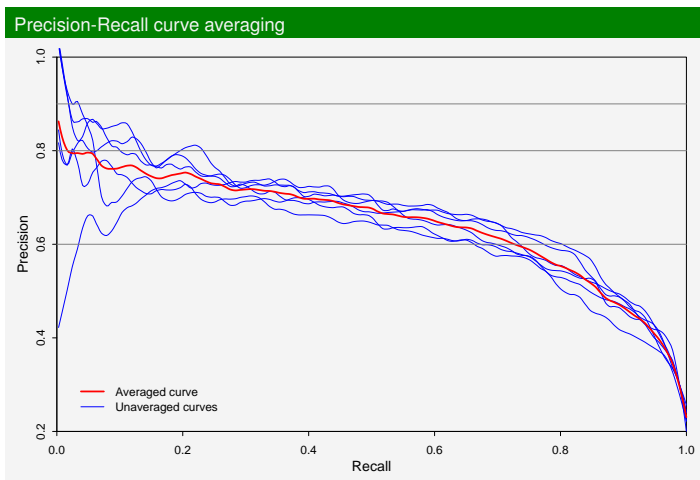
- ▶ graphical plots of **recall** vs. **precision**
- ▶ the closer to the top and right, the better ranking performance
- ▶ estimated for each **eval fold** and **vertically averaging**



## Visual evaluation: Precision-Recall curves

15/30

- ▶ 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**

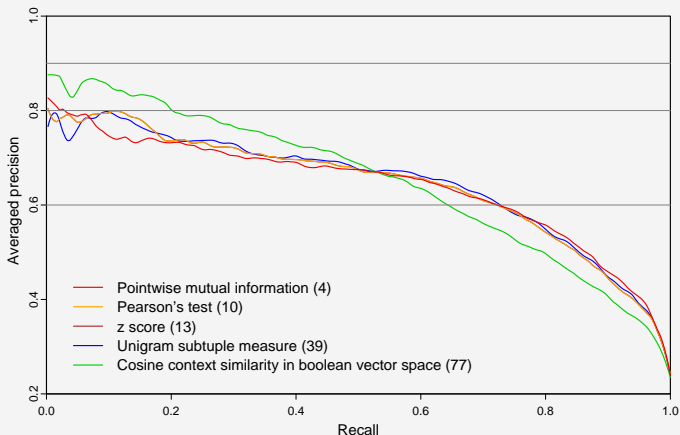




## Evaluation results: Precision-Recall curves

16/30

## The best-performing association measures



## Evaluation measure: Average Precision

17/30

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

## Evaluation measure: Average Precision

17/30

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

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

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

17/30

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

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

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

89.6% = AP

## Evaluation measure: Average Precision

17/30

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

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

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

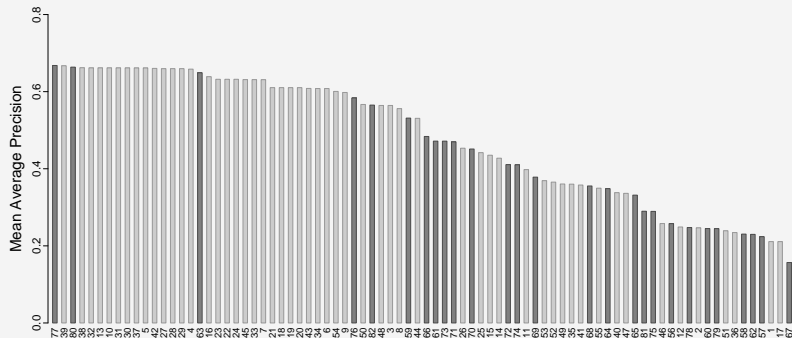
3) **Mean Average Precision:**  $E[AP]$   $MAP = \frac{1}{6} \sum_{i=1}^6 AP_i$

89.6% = AP

## Overall results: Mean Average Precision

18/30

MAP of all lexical association measures in descending order

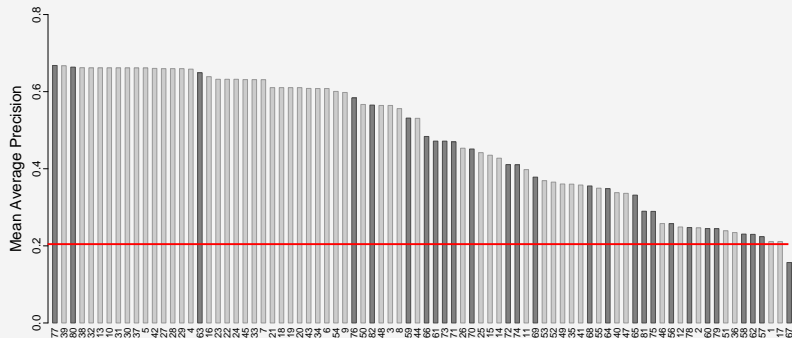


- ▶ Baseline (*ratio of true collocations*): 21.02%
- ▶ Best context-based measure (■): Cosine similarity in vector space: 66.79%
- ▶ Best statistical association measure (■): Unigram subtuple measure: 66.72%
- ▶ Best 16 measures – statistically indistinguishable MAP ~ current state of the art

## Overall results: Mean Average Precision

18/30

MAP of all lexical association measures in descending order

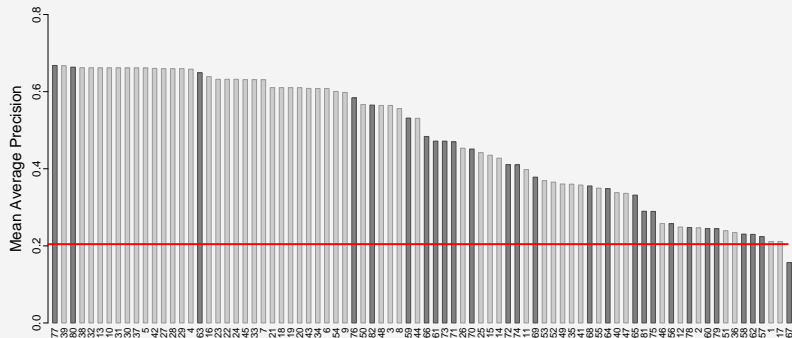


- ▶ Baseline (*ratio of true collocations*): **21.02%**
- ▶ Best context-based measure (■): Cosine similarity in vector space: **66.79%**
- ▶ Best statistical association measure (■): Unigram subtuple measure: **66.72%**
- ▶ Best 16 measures – statistically indistinguishable MAP ~ current **state of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

## Overall results: Mean Average Precision

18/30

MAP of all lexical association measures in descending order



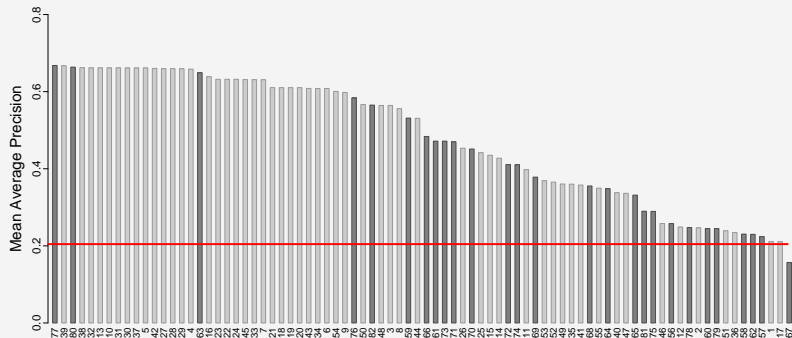
- ▶ Baseline (*ratio of true collocations*): 21.02%
- ▶ Best context-based measure (■): Cosine similarity in vector space: 66.79%
- ▶ Best statistical association measure (■): Unigram subtuple measure: 66.72%
- ▶ Best 16 measures – statistically indistinguishable MAP ~ current state of the art



## Overall results: Mean Average Precision

18/30

MAP of all lexical association measures in descending order

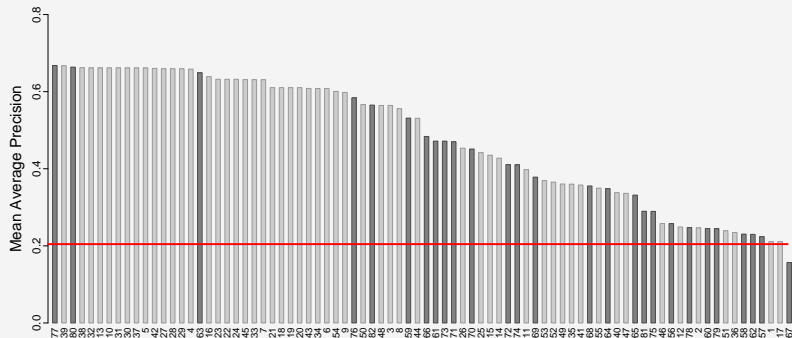


- ▶ Baseline (*ratio of true collocations*): 21.02%
- ▶ Best context-based measure (■): Cosine similarity in vector space: 66.79%
- ▶ Best statistical association measure (■): Unigram subtuple measure: 66.72%
- ▶ Best 16 measures – statistically indistinguishable MAP ~ current state of the art

## Overall results: Mean Average Precision

18/30

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

19/30

### Motivation

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

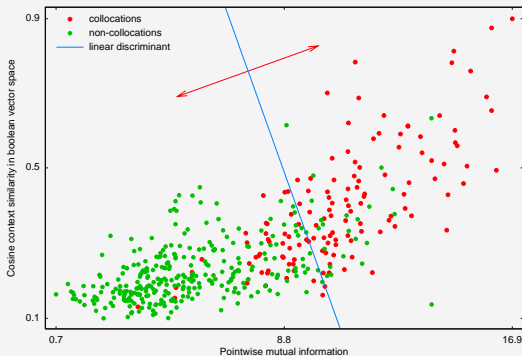
# Combining association measures

19/30

## Motivation

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

5% data sample from PDT-Dep

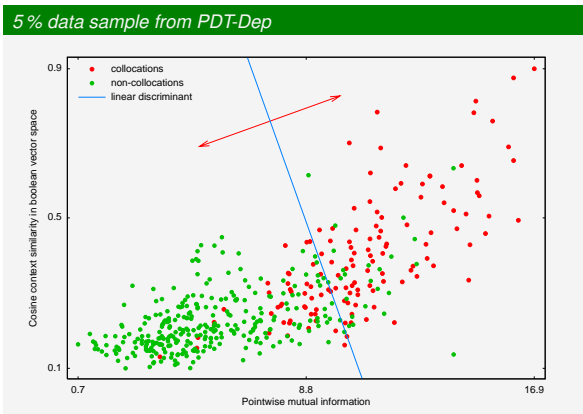


## Combining association measures

19/30

### Motivation

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



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

## Combination models

20/30

### Framework

- ▶ each collocation candidate  $\mathbf{x}^i$  is described by the **feature vector**  $\mathbf{x}^i = (x_1^i, \dots, x_{82}^i)^T$  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  $\mathbf{x}^i$
- ▶ e.g. **linear combination** of association scores:  $f(\mathbf{x}^i) = w_0 + w_1 x_1^i + \dots + w_{82} x_{82}^i$

### 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

20/30

### Framework

- ▶ each collocation candidate  $\mathbf{x}^i$  is described by the **feature vector**  $\mathbf{x}^i = (x_1^i, \dots, x_{82}^i)^T$  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  $\mathbf{x}^i$
- ▶ e.g. **linear combination** of association scores:  $f(\mathbf{x}^i) = w_0 + w_1 x_1^i + \dots + w_{82} x_{82}^i$

### 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  $\mathbf{x}^i$  is described by the **feature vector**  $\mathbf{x}^i = (x_1^i, \dots, x_{82}^i)^T$  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  $\mathbf{x}^i$
- ▶ e.g. **linear combination** of association scores:  $f(\mathbf{x}^i) = w_0 + w_1 x_1^i + \dots + w_{82} x_{82}^i$

### 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

21/30

## 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**



## Results: Mean Average Precision

<i>method</i>	<i>MAP</i>	<i>+%</i>
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

21/30

## 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**



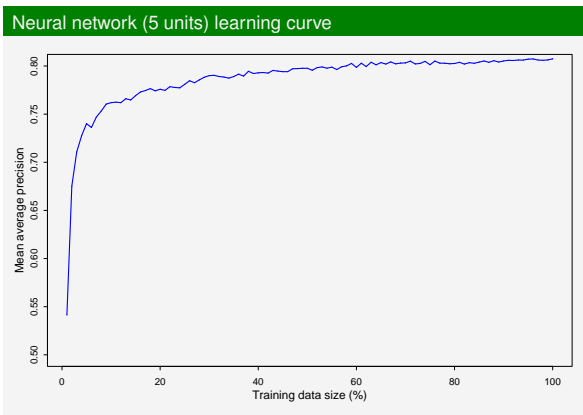
## Results: Mean Average Precision

<i>method</i>	<i>MAP</i>	<i>+%</i>
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



# Learning curve analysis

23/30



- ▶ 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

24/30

### 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

<i>method</i>	<i>MAP</i>	<i>+%</i>
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

24/30

### 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

<i>method</i>	<i>MAP</i>	<i>+%</i>
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

24/30

### 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

<i>method</i>	<i>MAP</i>	<i>+%</i>
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

25/30

## Motivation

- ▶ “Ocamo’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

25/30

## Motivation

- ▶ “Ockham’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

25/30

### 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

26/30

### Agglomerative hierarchical clustering

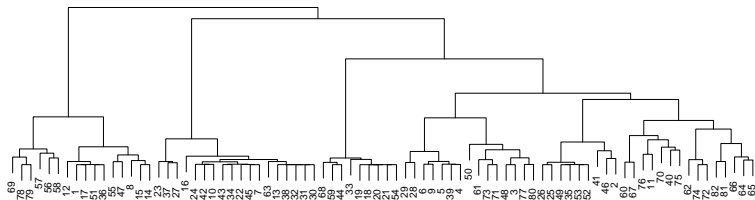
- ▶ groups the measures with the same/similar contribution to the model
- ▶ begins with each measure as a separate cluster and merge them into successively larger clusters
- ▶ distance metrics =  $1 - |\textit{Pearson's correlation}|$  (estimated on the *held-out* fold)

## Model reduction: 1) Clustering

26/30

### 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 =  $1 - |\text{Pearson's correlation}|$  (estimated on the *held-out* fold)

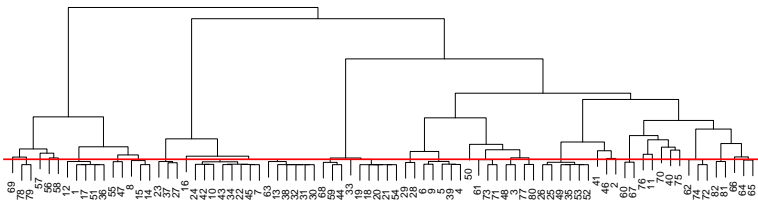


## Model reduction: 1) Clustering

26/30

### 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 =  $1 - |\text{Pearson's correlation}|$  (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

27/30

### Iterative procedure

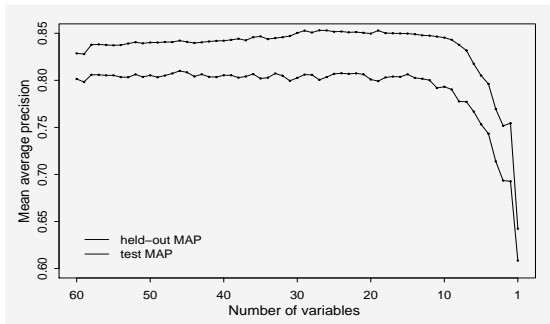
- ▶ initiated with the 60 variables/measures
- ▶ in each iteration we remove the variable causing minimal performance degradation when not used in the model (by MAP on the *held-out* fold)
- ▶ stops before the degradation becomes statistically significant

## Model reduction: 2) Stepwise variable removal

27/30

### 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

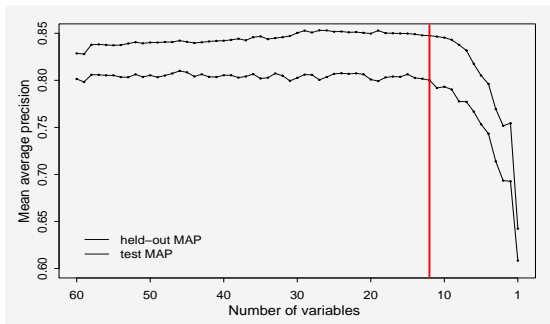


## Model reduction: 2) Stepwise variable removal

27/30

### 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



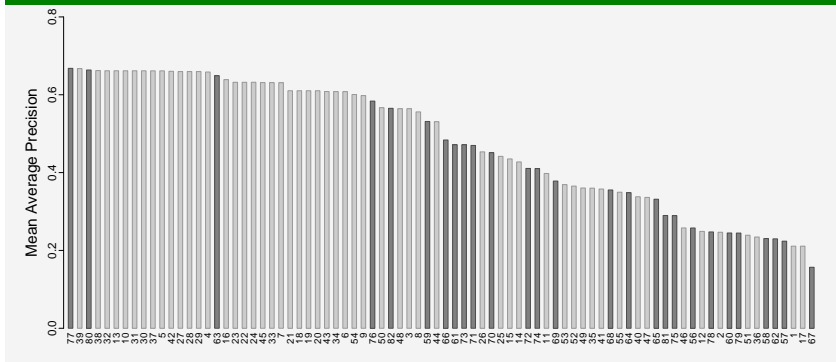
- ▶ the final model contains 13 variables/lexical association measures



## Model reduction: Process overview

28/30

MAP of individual lexical association measures

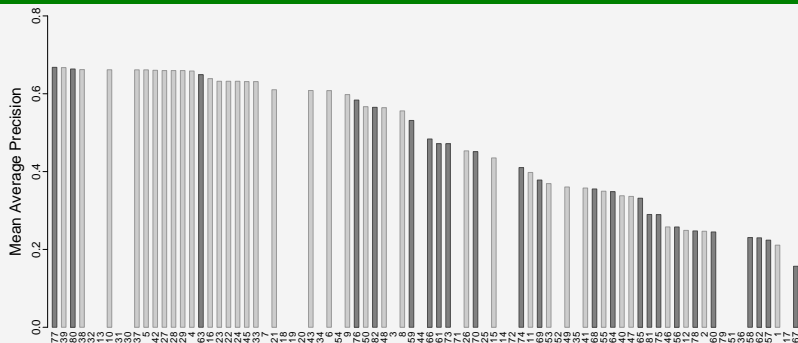


- ▶ procedure initiated with all **82** association measures
- ▶ highly correlated measures removed in the first phase (*clustering*)
- ▶ **13** measures left after the second phase (*stepwise removal*)
  - ≡ 4 statistical association measures (■) + 9 context-based measures (■)

## Model reduction: Process overview

28/30

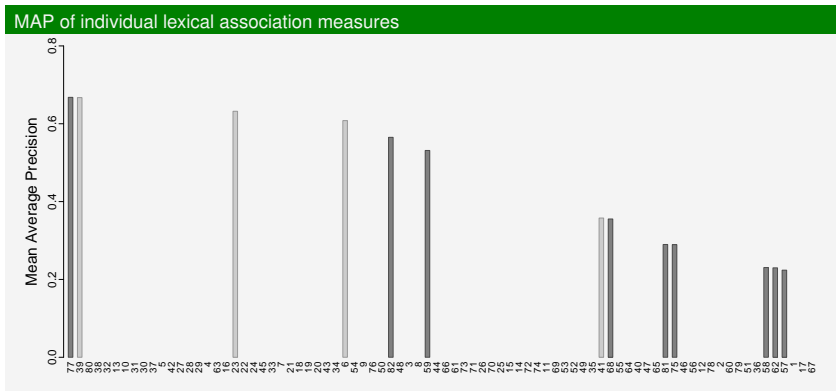
MAP of individual lexical association measures



- ▶ 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

28/30

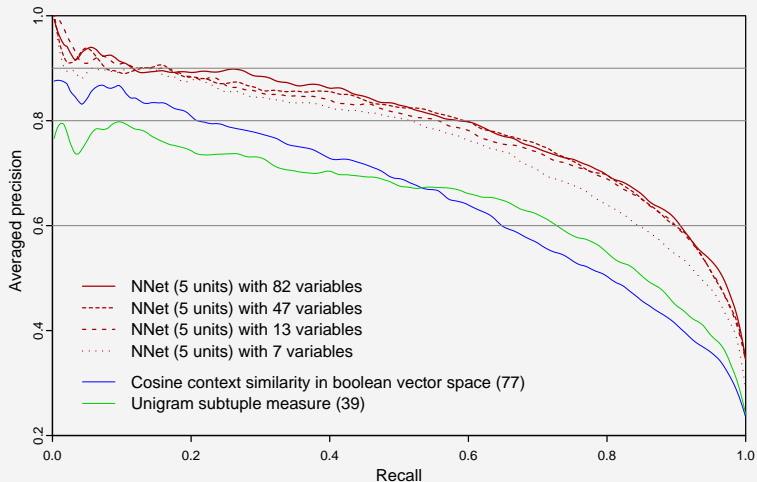


- ▶ procedure initiated with all **82** association measures
- ▶ highly correlated measures removed in the first phase (*clustering*)
- ▶ **13** measures left after the second phase (*stepwise removal*)
  - = 4 statistical association measures (■) + 9 context-based measures (■)

## Model reduction results: Precision-Recall curves

29/30

Reduced combination models compared with the best association measures



## 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 (*average precision, crossvalidation, significance tests*)
- ▶ reference data sets used in MWE 2008 Shared Task

# Conclusions

30/30





## 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 (*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*, Springer, 2009 (accepted).
-  Pavel Pecina: **Lexical Association Measures: Collocation Extraction**, *PhD Thesis*, Charles University, Prague, Czech Republic, 2008.
-  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.

## 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, morphologically 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

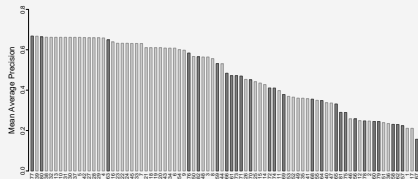


## Reference data summary

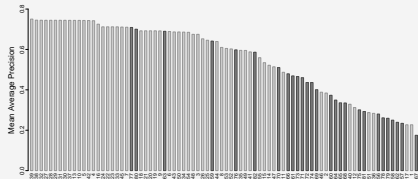
<i>reference data set</i>	<i>PDT-Dep</i>	<i>PDT-Surf</i>	<i>CNC-Surf</i>	<i>PAR-Dist</i>
source corpus	<i>PDT</i>	<i>PDT</i>	<i>CNC</i>	<i>PAROLE</i>
language	<i>Czech</i>	<i>Czech</i>	<i>Czech</i>	<i>Swedish</i>
morphology	<i>manual</i>	<i>manual</i>	<i>auto</i>	<i>auto</i>
syntax	<i>manual</i>	<i>none</i>	<i>none</i>	<i>none</i>
bigram types	<i>dependency</i>	<i>surface</i>	<i>surface</i>	<i>distance</i>
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	<b>12 232</b>	<b>10 021</b>	<b>9 868</b>	<b>17 027</b>
data sample size	100 %	100 %	0.66 %	1.90 %
true collocations	<b>2 557</b>	<b>2 293</b>	<b>2 263</b>	<b>1 292</b>
baseline precision (%)	<b>21.02</b>	<b>22.88</b>	<b>22.66</b>	<b>7.59</b>

## Context-based vs. statistical association measures

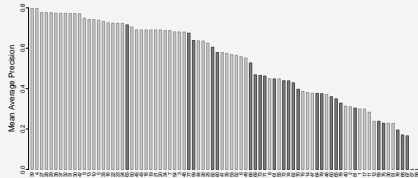
PDT-Dep



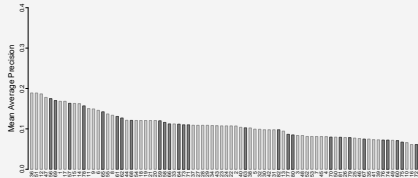
PDT-Surf



CNC-Surf

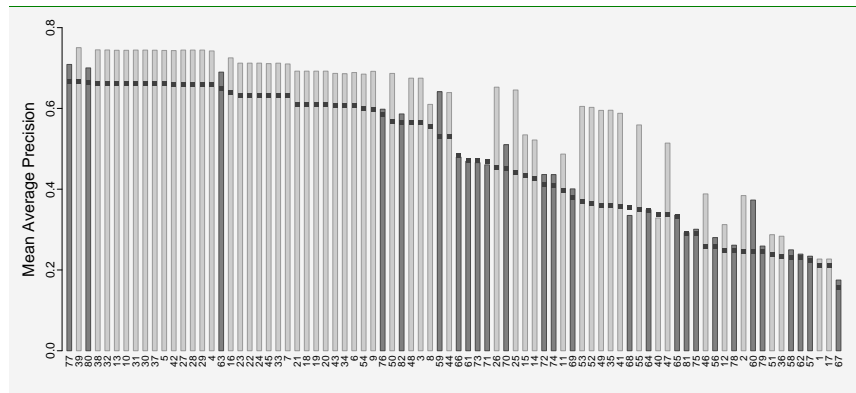


PAR-Dist



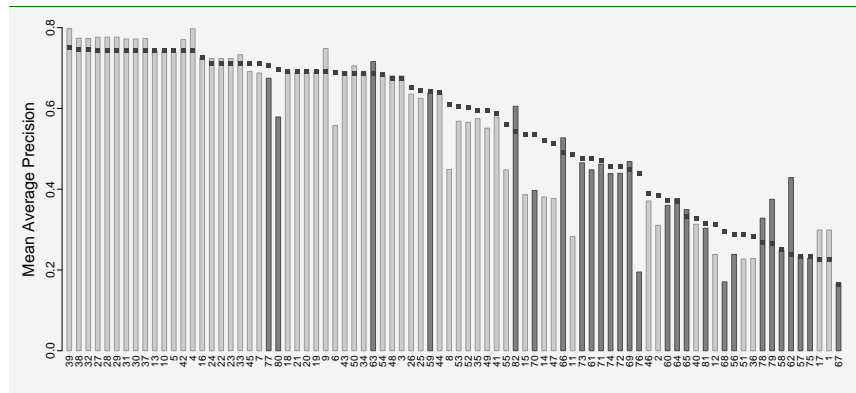
Results / Mean average precision: *PDT-Dep* vs. *PDT-Surf*

## Dependency bigrams vs. surface bigrams



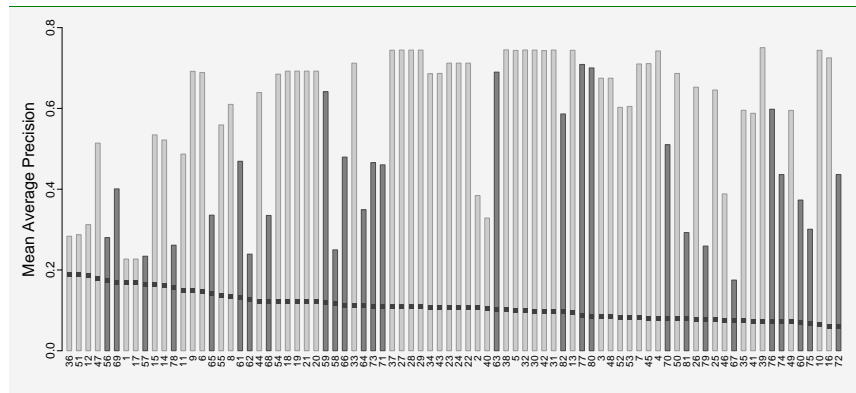
Results / Mean average precision: *PDT-Surf* vs. *CNC-Surf*

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

