Automatic Collocation Extraction from Text Corpora

Pavel Pecina

Ústav formální a aplikované lingvistiky
MFF UK Praha

May 17, 2004
Outline

1. The notion of collocation
   - Motivation
   - Few definitions
   - Characteristic features, classification and categorization

2. Methodology of collocation extraction
   - Phrase Extraction
   - Collocation identification

3. Experiments
   - Toolkit
   - Data
   - Basic Methods Evaluation
   - Advanced Methods

4. Summary
   - Conclusion, Future work, Used Tools
Outline

1. The notion of collocation
   - Motivation
   - Few definitions
   - Characteristic features, classification and categorization

2. Methodology of collocation extraction
   - Phrase Extraction
   - Collocation identification

3. Experiments
   - Toolkit
   - Data
   - Basic Methods Evaluation
   - Advanced Methods

4. Summary
   - Conclusion, Future work, Used Tools
Well known problems

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

Machine translation
- *Where to brake a sentence into chunks?*
  She likes ice cream pancakes.

Information retrieval
- *Which multiword terms to index?*
  Our new friend is from New York.

Word sense disambiguation
- *How to distinguish between possible word senses?*
  My uncle owns a wine yard.
Other well known problems

spell/grammar/style-checking
- *Is this text written correctly?*
  Meals will be served outside, weather allowing.

text classification and summarization
- *What is this text about?*
  Carriage return is necessary here.

Language modeling (text/speech synthesis)
- *How to create a fluent sentence?*
  Could you hand me salt and pepper?

Corpus-based language teaching/learning
- *What kinds of multiword expressions to teach?*
  When she kicked his head he kicked the bucket.
What are we looking for?

- noun phrases: disk drive, weapons of mass destruction
- light verbs compounds: keep an eye, make a decision
- phrasal verbs: make up, give up, tell off
- stock phrases: bacon and eggs, salt and pepper
- idioms: hear it through the grapevine
- technological expressions: object oriented language
- proper names: Joe Black, Prague Spring
- frequent usages: game over, good morning
- multiword units w/ independent existence: white wine, Far East
- close associations between words: knock on a door, thick hair
What are we looking for?

- noun phrases: disk drive, weapons of mass destruction
- light verbs compounds: keep an eye, make a decision
- phrasal verbs: make up, give up, tell off
- stock phrases: bacon and eggs, salt and pepper
- idioms: hear it through the grapevine
- technological expressions: object oriented language
- proper names: Joe Black, Prague Spring
- frequent usages: game over, good morning
- multiword units w/ independent existence: white wine, Far East
- close associations between words: knock on a door, thick hair

Collocations.
Definitions ...

Firth (1951)

"Collocations of a given word are statements of the habitual or customary places of that word."

Choueka (1988)

"A collocation is a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components."
Manning (1999)

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

Radev (1998)

“A collocation is a group of words that occur together more often than by a chance.”
“A collocation is an expression consisting of two or more words that form a grammatical and semantic unit.”
Characteristic Features

Non-compositionality
    kick the bucket, carriage return, white man

Non-substituability
    yellow wine, hit the bucket, make homework

Non-modifiability
    give a small hand, poor as a church mice

Not straightforward translation
    ice cream, to be right

Domain-dependency
    carriage return,

“Subjectivity”
    game over, new company
Classification

Semantics
- compositional, noncompositional

Consecutivity
- free, fixed

Functionality
- idioms, proper names, technical terms, phrasal verbs, light verbs

Word usage
- A→N, N→A, D→V, R→N
**Grammar Patterns**

<table>
<thead>
<tr>
<th>Part-Of-Speech</th>
<th>Line</th>
<th>Dependency Types</th>
</tr>
</thead>
<tbody>
<tr>
<td>A N</td>
<td>lineární funkce</td>
<td>Atr cenný papír</td>
</tr>
<tr>
<td>N N</td>
<td>následník trůnu</td>
<td>Sb soud rozhodl</td>
</tr>
<tr>
<td>D A N</td>
<td>objektově orientovaný jazyk</td>
<td>Obj dávat přednost</td>
</tr>
<tr>
<td>N A N</td>
<td>zbraně hromadného ničení</td>
<td>Adv zdravotně postižený</td>
</tr>
<tr>
<td>V R N</td>
<td>přijít k sobě</td>
<td></td>
</tr>
</tbody>
</table>
Outline

1. The notion of collocation
   - Motivation
   - Few definitions
   - Characteristic features, classification and categorization

2. Methodology of collocation extraction
   - Phrase Extraction
   - Collocation identification

3. Experiments
   - Toolkit
   - Data
   - Basic Methods Evaluation
   - Advanced Methods

4. Summary
   - Conclusion, Future work, Used Tools
Phrase extraction

1. extracting all possible candidates for collocations
   - consequent word n-grams
   - sliding window
   - syntactical subtrees

2. collecting their occurrence statistics
   - contingency tables
   - empirical context
### Contingency table: observed frequencies

**bigram: \( xy \)**

<table>
<thead>
<tr>
<th></th>
<th>( X=x )</th>
<th>( X \neq x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y=y )</td>
<td>( O_{11} )</td>
<td>( O_{12} )</td>
</tr>
<tr>
<td>( Y \neq y )</td>
<td>( O_{21} )</td>
<td>( O_{22} )</td>
</tr>
<tr>
<td></td>
<td>( C_1 )</td>
<td>( C_2 )</td>
</tr>
</tbody>
</table>

#### Example: černý trh

<table>
<thead>
<tr>
<th></th>
<th>( X=černý )</th>
<th>( X \neq černý )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y=trh )</td>
<td>černý trh</td>
<td>domácí trh</td>
</tr>
<tr>
<td>( Y \neq trh )</td>
<td>černý čaj</td>
<td>zelený čaj</td>
</tr>
</tbody>
</table>
The notion of collocation
Methodology of collocation extraction
Experiments
Summary

Phrase Extraction
Collocation identification

Contingency table: observed frequencies

<table>
<thead>
<tr>
<th>bigram: xy</th>
<th>X=x</th>
<th>X≠x</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y=y</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Y≠y</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

example: černý trh

<table>
<thead>
<tr>
<th></th>
<th>X=černý</th>
<th>X≠černý</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y=trh</td>
<td>černý trh</td>
<td>domácí trh</td>
</tr>
<tr>
<td>Y≠trh</td>
<td>černý čaj</td>
<td>zelený čaj</td>
</tr>
</tbody>
</table>
Average Word Context

Example

zlepšení situace. Kapitálový trh je však stále nelikvidní
že to není samostatný trh a že je součástá širšího
bariérách v přístupu na trh, cenových rozdílech,
banky. Americký akciový trh byl za silného obchodování
jít se svou kuží na trh. Pro vydání mluvila zejména

Context word probability distribution $P(w_i|x)$
Few different basic approaches

1. Cooccurrence statistics
2. Hypothesis tests
3. Association estimation
4. Information theory measures
5. Context similarity measures
Cooccurrence statistics

- Joint probability $P(xy)$
- Conditional probability $P(y|x)$
- Reverse conditional probability $P(y|x)$
- Symmetric conditional probability $P(y|x)P(x|y)$
### Hypothesis testing:

Null hypothesis: word occurrences are independent

\[ H_0 : P(xy) = P(x)P(y) \]

#### bigram: \( xy \)

<table>
<thead>
<tr>
<th></th>
<th>( X=x )</th>
<th>( X \neq x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y=y )</td>
<td>( E_{11} = \frac{R_1C_1}{N} )</td>
<td>( E_{12} = \frac{R_1C_2}{N} )</td>
</tr>
<tr>
<td>( Y \neq y )</td>
<td>( E_{21} = \frac{R_2C_1}{N} )</td>
<td>( E_{22} = \frac{R_2C_2}{N} )</td>
</tr>
</tbody>
</table>
Hypothesis testing cont.

- z-score \( \frac{O_{11} - E_{11}}{\sqrt{E_{11}}} \)
- t-score \( \frac{O_{11} - E_{11}}{\sqrt{O_{11}}} \)
- \( \chi^2 \) score \( \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \)
- log-likelihood \( 2 \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}} \)
Association estimation

- Russel-Rao \( \frac{a}{a+b+c+d} \)
- Sokal-Michiner \( \frac{a+d}{a+b+c+d} \)
- Rogers-Tanimoto \( \frac{a+d}{a+2b+2c+d} \)
- Hamann \( \frac{(a+d)-(b+c)}{a+b+c+d} \)
- Sokal-Sneath 3\(^{rd}\) \( \frac{b+c}{a+d} \)
- Jaccard \( \frac{a}{a+b+c} \)
- Kulczynski 1\(^{st}\) \( \frac{a}{b+c} \)
- Sokal-Sneath 2\(^{th}\) \( \frac{a}{a+2(b+c)} \)
- Kulczynski 2\(^{nd}\) \( \frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c}) \)
Information theory and context similarity measures

- pointwise mutual information $\log \frac{P(xy)}{P(x)P(y)}$
- local mutual information $NP(xy) \log \frac{P(xy)}{P(x)P(y)}$

- Cross Entropy $- \sum_{w \in C} P(w|x) \log_2 P(w|y)$
- Intersection measure $\frac{2|C_x \cap C_y|}{|C_x| + |C_y|}$
- Euclidean norm $\sqrt{\sum_{w \in C} (P(w|x) - P(w|y))^2}$
- Cosine norm $\frac{\sum_{w \in C} P(w|x)P(w|y)}{\sum_{w \in C} P(w|x)^2 \sum_{w \in C} P(w|y)^2}$
- L1 norm $\sum_{w \in C} |P(w|x) - P(w|y)|$
Outline

1. The notion of collocation
   - Motivation
   - Few definitions
   - Characteristic features, classification and categorization

2. Methodology of collocation extraction
   - Phrase Extraction
   - Collocation identification

3. Experiments
   - Toolkit
   - Data
   - Basic Methods Evaluation
   - Advanced Methods

4. Summary
   - Conclusion, Future work, Used Tools
Task setup

1. Implementation of toolkit for statistical analysis of word cooccurrences
2. Collecting of basic methods for collocation extraction

1. Implementation of the basic methods
2. Evaluation of the basic methods
3. Experiments with advanced methods
Toolkit

- fully functional prototype implementation in Perl
- *Input*: plain text/ morphological level/ analytical level
- *Output*: collocation candidates with values of all specified measures and scores
Word Base Forms

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

```
<nenahraditelná<l>nahraditelný_(*4)<t>AAFS1----1N----<r>8<g>7
↓ ↓ ↓ ↓↓
nahraditelný_(*4) A F 1N
↓
<f>nahraditelný_(*4)<t>A*F1N</f>
↓
nenahraditelná
```
Data

Prague Dependency Treebank

- base form types: 66 662
- bigram types: 306 845
- experiments performed on dependency bigrams with frequency $> 5$: 21 595
- all these collocation candidates manually evaluated ...
All dependency bigrams with frequency > 5 classified into 6 groups:

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Collocation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>kámen úrazu, slepá ulička, železná opona</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>bílý dum, černý trh, poslední slovo, pata kolmice</td>
<td>201</td>
</tr>
<tr>
<td>3</td>
<td>šifrovací klíč, atomová energie, Baník Ostrava</td>
<td>2460</td>
</tr>
<tr>
<td>2</td>
<td>dávat přednost, minulé století, starosta města</td>
<td>443</td>
</tr>
<tr>
<td>1</td>
<td>na Slovensko, do Portugalska</td>
<td>484</td>
</tr>
<tr>
<td>0</td>
<td>(non-collocations)</td>
<td>18002</td>
</tr>
</tbody>
</table>
All dependency bigrams with frequency > 5 classified into 6 groups:

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Bigrams</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>kámen úrazu, slepá ulička, železná opona</td>
<td>3595</td>
</tr>
<tr>
<td>4</td>
<td>bilý dum, černý trh, poslední slovo, pata kolmice</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>šifrovací klíč, atomová energie, Baník Ostrava</td>
<td>18002</td>
</tr>
<tr>
<td>2</td>
<td>dávat přednost, minulé století, starosta města</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>na Slovensko, do Portugalska</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>(non-collocations)</td>
<td></td>
</tr>
</tbody>
</table>
### Evaluation Data

All dependency bigrams with frequency $> 5$ classified into 6 groups:

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Bigram</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>kámen úrazu, spleťa ulička, železná opona</td>
<td>2668</td>
</tr>
<tr>
<td>4</td>
<td>bílý dum, černý trh, poslední slovo, pata kolmice</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>šifrovací klíč, atomová energie, Baník Ostrava</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>dávat přednost, minulé století, starosta města</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>na Slovensko, do Portugalska</td>
<td>18929</td>
</tr>
<tr>
<td>0</td>
<td>(non-collocations)</td>
<td></td>
</tr>
</tbody>
</table>
Basic Methods

Pattern filtering
- Part of Speech pattern
- Dependency pattern

Association measures and scores
- Cooccurrence statistics
- Likelihood measures
- Hypothesis testing
- Association estimation
- Information theory measures
- Context similarity measures
Evaluation: trading recall for precision

Precision

\[ P = \frac{\# \text{selected collocations}}{\# \text{selected bigrams}} \in <0, 1> \]

Recall

\[ R = \frac{\# \text{selected collocations}}{\# \text{all collocations}} \in <0, 1> \]
Recall and precision

Example: Mutual Information

\[ \log_2 \frac{P(xy)}{P(x)P(y)} \]
Recall and precision

Example: \( t \) score

\[
\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}
\]
Recall and precision

Example: $z$ score

\[
\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}
\]

Precision Recall 20: $Z$ score

Recall (%)  
0 20 40 60 80 100

Precision (%)  
0 20 40 60 80 100

Precision Recall 20: $Z$ score
**Evaluation results**

**Overview**

![Precision Recall Graph](image_url)

**Pavel Pecina**

*Automatic Collocation Extraction from Text Corpora*
Advanced Methods: motivation

Example: Mutual Information vs. Cosine context similarity

11: Mutual Information
85: Cosine similarity

Pavel Pecina
Automatic Collocation Extraction from Text Corpora
Advanced Methods: idea

Statistical learning problem

- for each bigram we get set of features (categories, scores etc.)
  \[ x_i = (x_1, x_2, \ldots, x_{90}) \]
- each bigram we want to classify as collocation or noncolloc.
  \[ f(x_i) = y_i, y_i = 0, 1 \]
- so we are looking for function that minimizes a risk functional
  \[ \min \sum_i Q(f(x_i), y_i) \]
Statistical learning problem

- but classification might be hard, what about regression?

\[ f(x_i) = y_y, y_i \in \langle 0, 1 \rangle \]

- Isn’t \( R^{90} \) too much? What about feature selection?

- Yes! And how to do it?
  - Liner discriminant
  - General linear models - logistic regression
  - Neural networks
  - Support vector Machines
The notion of collocation
Methodology of collocation extraction
Experiments
Summary
Toolkit
Data
Basic Methods Evaluation
Advanced Methods

Result

Support Vector Machines

Precision (%)
Recall (%)
Outline

1. The notion of collocation
   - Motivation
   - Few definitions
   - Characteristic features, classification and categorization

2. Methodology of collocation extraction
   - Phrase Extraction
   - Collocation identification

3. Experiments
   - Toolkit
   - Data
   - Basic Methods Evaluation
   - Advanced Methods

4. Summary
   - Conclusion, Future work, Used Tools
Conclusion

Achieved results

- implementation and evaluation of basic methods for collocation extraction
- promising results with advanced methods

Future work

- experiments with advanced methods
- evaluation of advanced methods
- experiments on English data
Used tools and toolkits

**R-project**
- a language and environment for statistical computing and graphics
- extremely powerful
- GNU GPL license
- [www.r-project.org](http://www.r-project.org)

**Torch**
- machine learning library
- C++, BSD license
- [www.torch.ch](http://www.torch.ch)