Automatic Collocation Extraction from Text Corpora

Pavel Pecina

Ústav formální a aplikovaňe lingvistiky
MFF UK Praha

May 17, 2004

Outline

Outline

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."

Other Definitions ...

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 that occur together more often than by a chance."

... and The Definition

"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

- $A \rightarrow N$, $N \rightarrow A$, $D \rightarrow V$, $R \rightarrow N$

Grammar Patterns

Part-Of-Speech

A N	lineární funkce
N N	následník trůnu
D A N	objektově orientovaný jazyk
N A N	zbraně hromadného ničení
V R N	přijít k sobě

Dependency Types

Atr	cenný papír
Sb	soud rozhodl
Obj	dávat přednost
Adv	zdravotně postižený

Outline

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*

	$X{=}x$	$X{\neq}x$	
$Y{=}y$	O_{11}	O_{12}	R_1
$Y{\neq}y$	O_{21}	O_{22}	R_2
	C_1	C_2	N

example: černý trh

	$X{=}$černý	$X{\neq}$černý	
$Y{=}$trh	černý trh	domácí trh	
$Y{\neq}$trh	černý čaj	zelený čaj	

Contingency table: observed frequencies

bigram: *xy*

	$X{=}x$	$X{\neq}x$	
$Y{=}y$	a	b	
$Y{\neq}y$	c	d	

example: černý trh

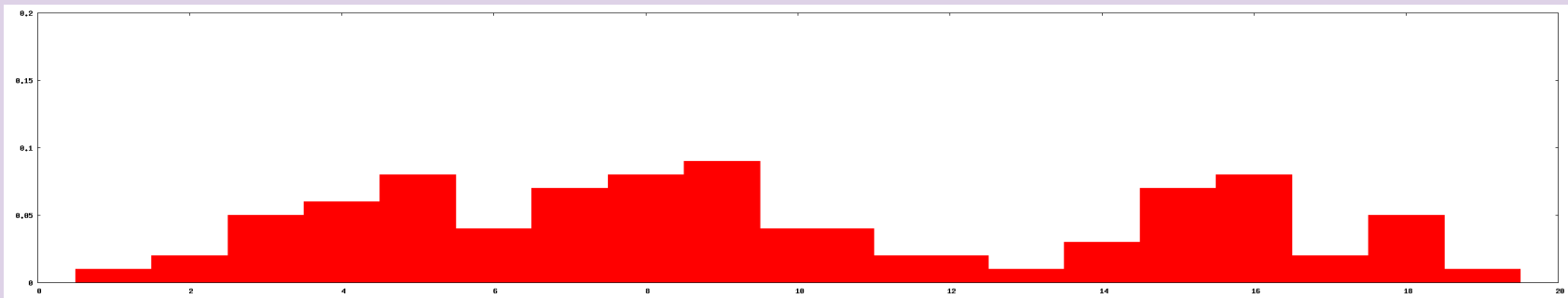
	$X{=}$*černý*	$X{\neq}$*černý*	
$Y{=}$*trh*	*černý trh*	*domácí trh*	
$Y{\neq}$*trh*	*černý čaj*	*zelený čaj*	

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 vydán i mluvila zejména

Context word probability distribution $P(w_i|x)$



Collocation Identification

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 $P(y|x)$
- Symetric conditional probability $P(y|x)P(x|y)$

Hypothesis testing:

Null hypothesis: word occurences are independent

$$\mathbf{H_0} : P(xy) = P(x)P(y)$$

bigram: *xy*

	$X{=}x$	$X{\neq}x$
$Y{=}y$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$Y{\neq}y$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

Hypothesis testing cont.

- z-score $\frac{O_{11}-E_{11}}{\sqrt{E_{11}}}$
- t-score $\frac{O_{11}-E_{11}}{\sqrt{O_{11}}}$
- χ^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 thory and ontext 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

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

<f>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 ...

Evaluation Data

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

5	kámen úrazu, slepá ulička, železná opona	7
4	bilý dum, černý trh, poslední slovo, pata kolmice	201
3	šifrovací klič, atomová energie, Baník Ostrava	2460
2	dávat přednost, minulé století, starosta města	443
1	na Slovensko, do Portugalska	484
0	(non-collocations)	18002

Evaluation Data

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

5	kámen úrazu, slepá ulička, železná opona	
4	bilý dum, černý trh, poslední slovo, pata kolmice	
3	šifrovací klič, atomová energie, Baník Ostrava	3595
2	dávat přednost, minulé století, starosta města	
1	na Slovensko, do Portugalska	
0	(non-collocations)	18002

Evaluation Data

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

5 4 3	kámen úrazu, slepá ulička, železná opona bilý dum, černý trh, poslední slovo, pata kolmice šifrovací klíč, atomová energie, Baník Ostrava	2668
2 1 0	dávat přednost, minulé století, starosta města na Slovensko, do Portugalska (non-collocations)	18929

Basic Methods

Pattern filtering

- Part of Speech pattern
- Dependecy 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

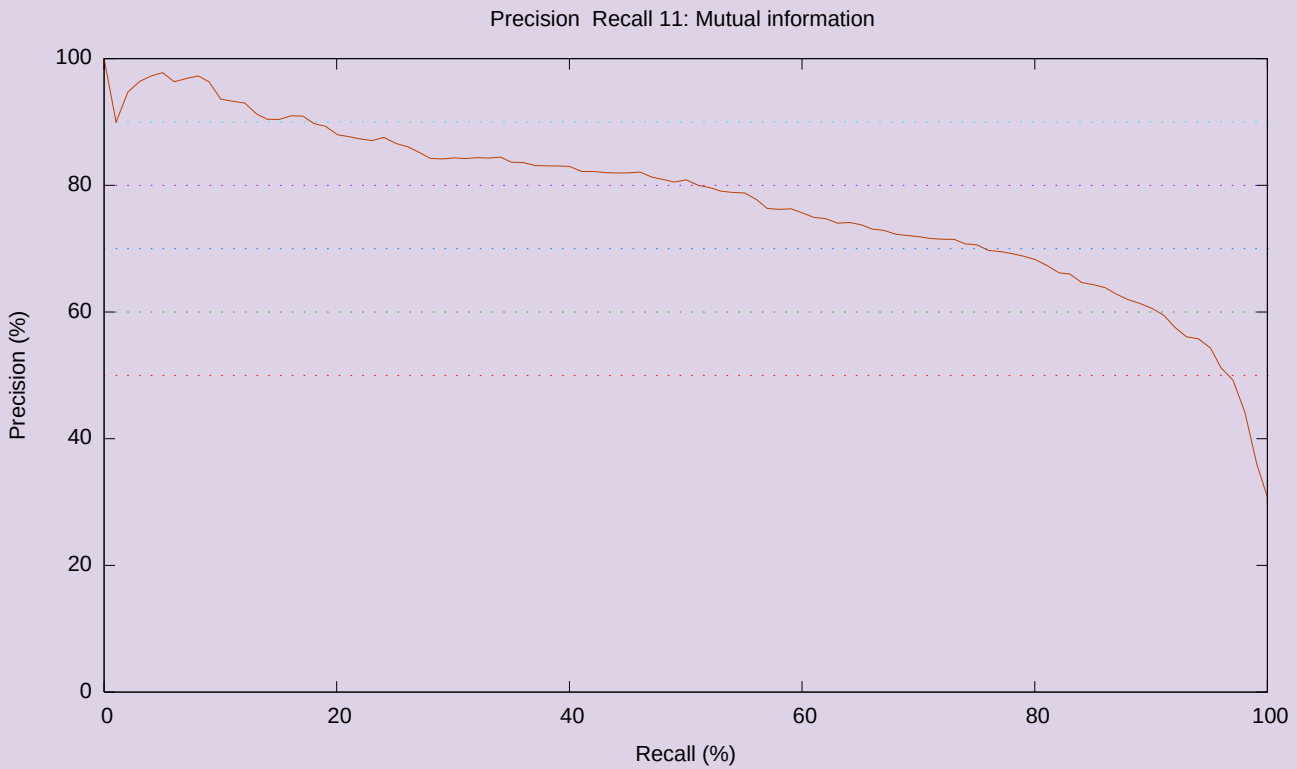
$$P = \frac{\#selected\ \ collocations}{\#selected\ \ bigrams} \in < 0, 1 >$$

Recall

$$R = \frac{\#selected\ \ collocations}{\#all\ \ collocation} \in < 0, 1 >$$

Recall and precision

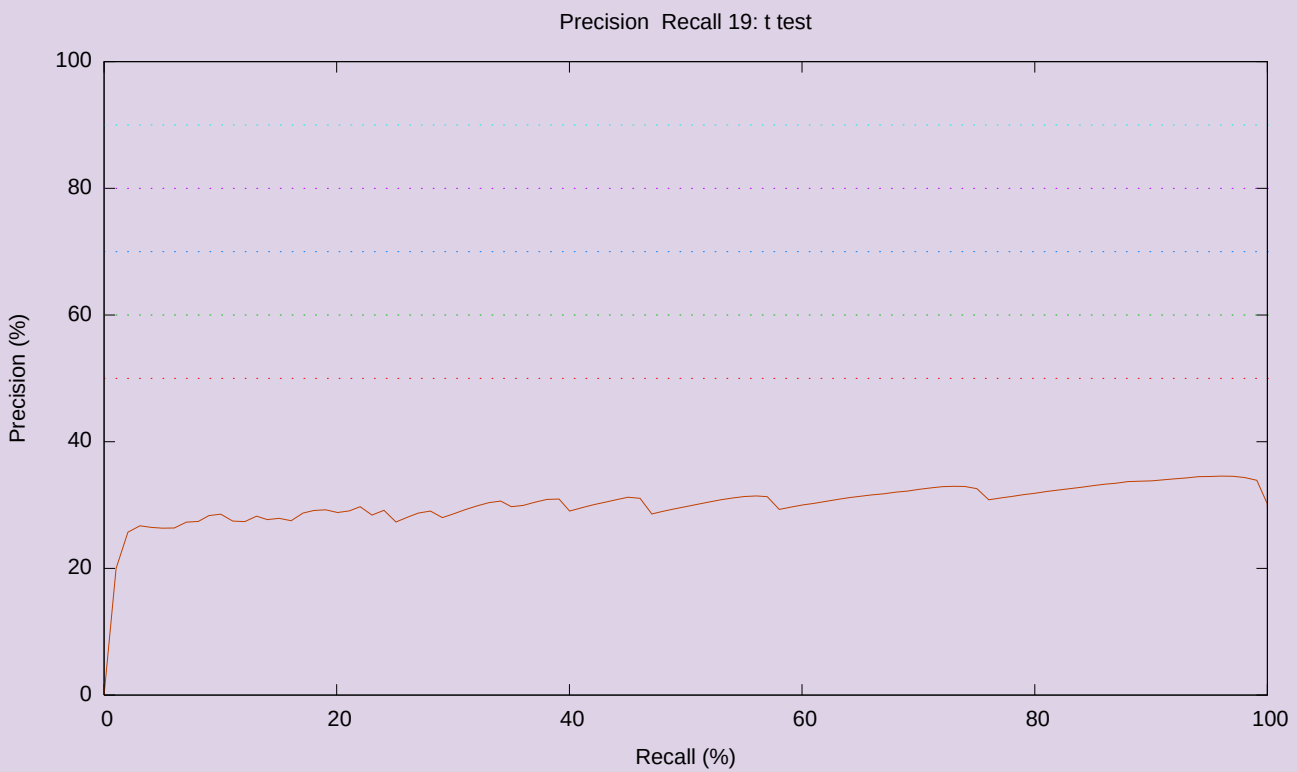
Example: Mutual Information



$$log_2 \frac{P(xy)}{P(x)P(y)}$$

Recall and precision

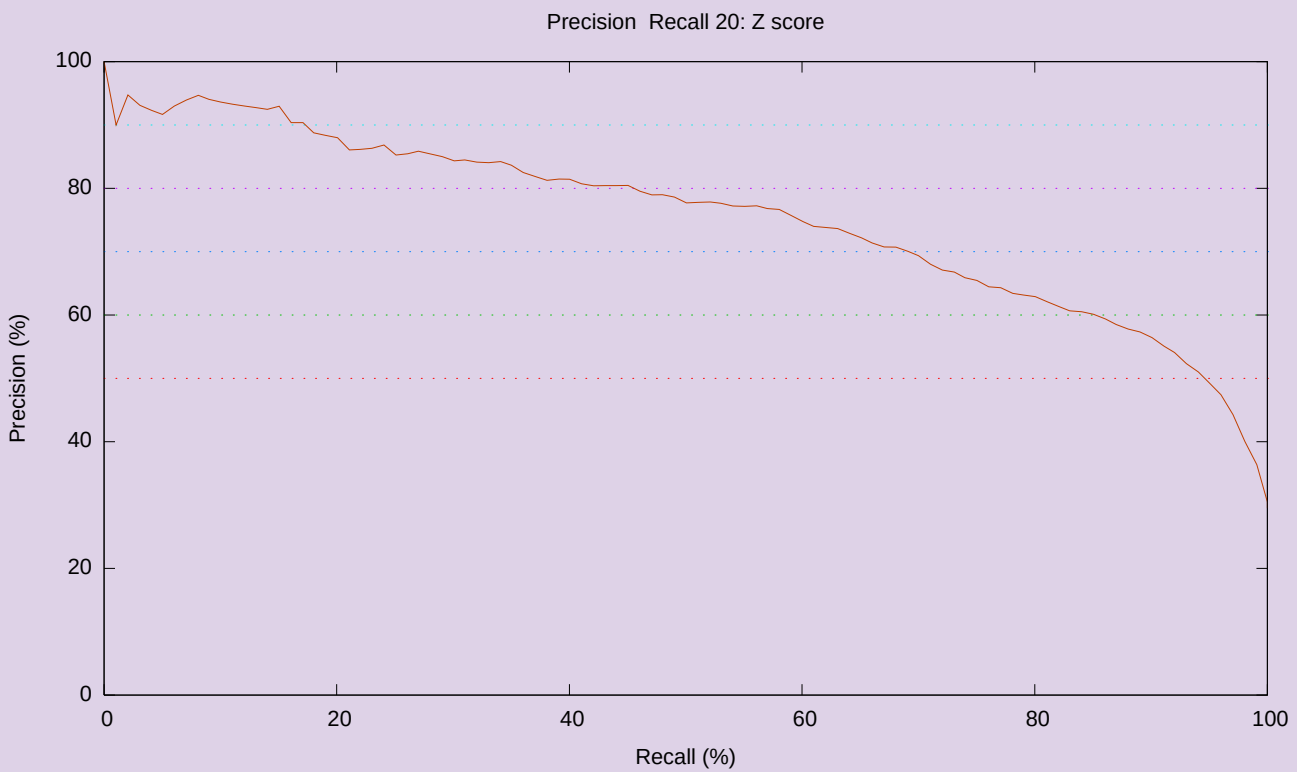
Example: t score



$$\frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$$

Recall and precision

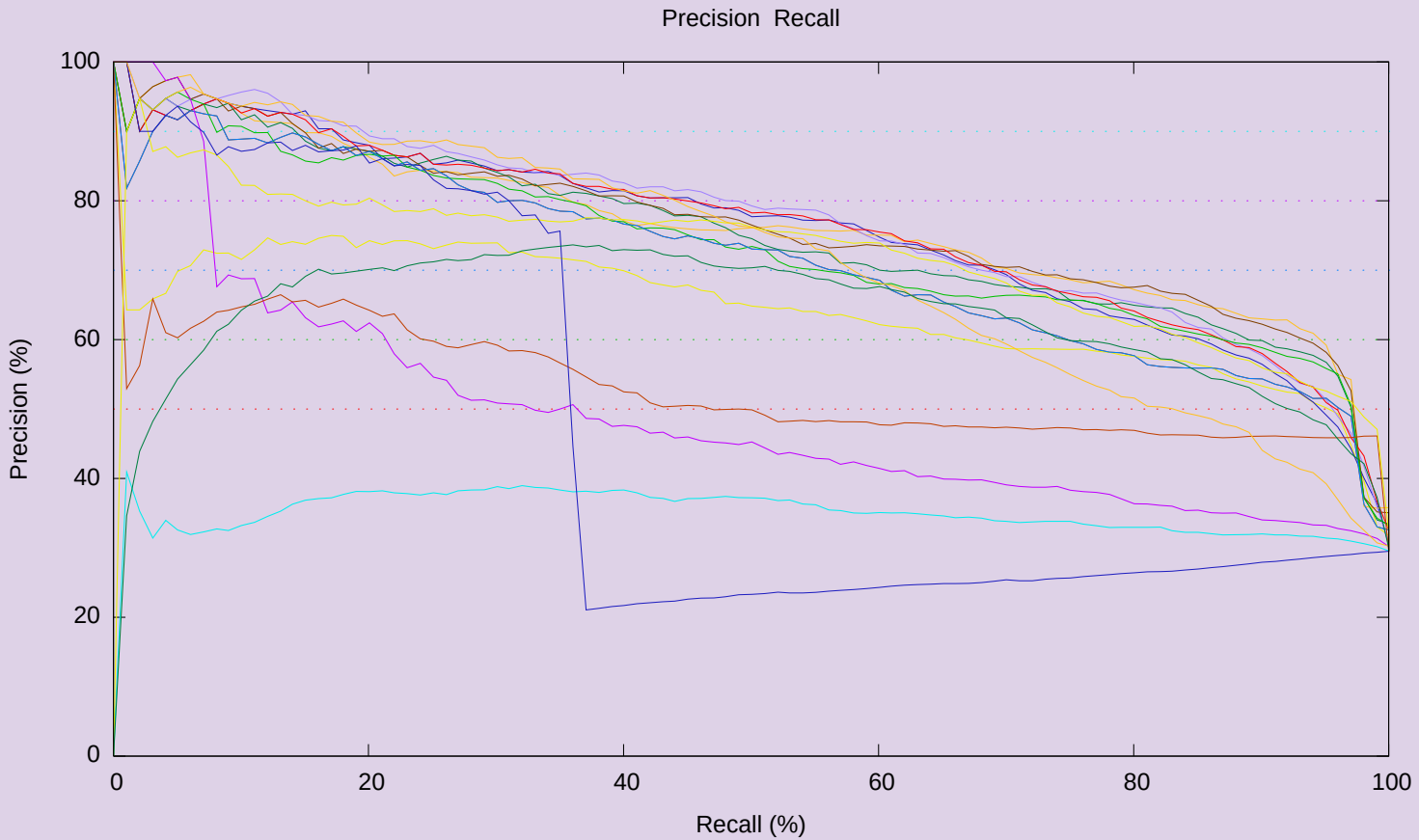
Example: z score



$$\frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

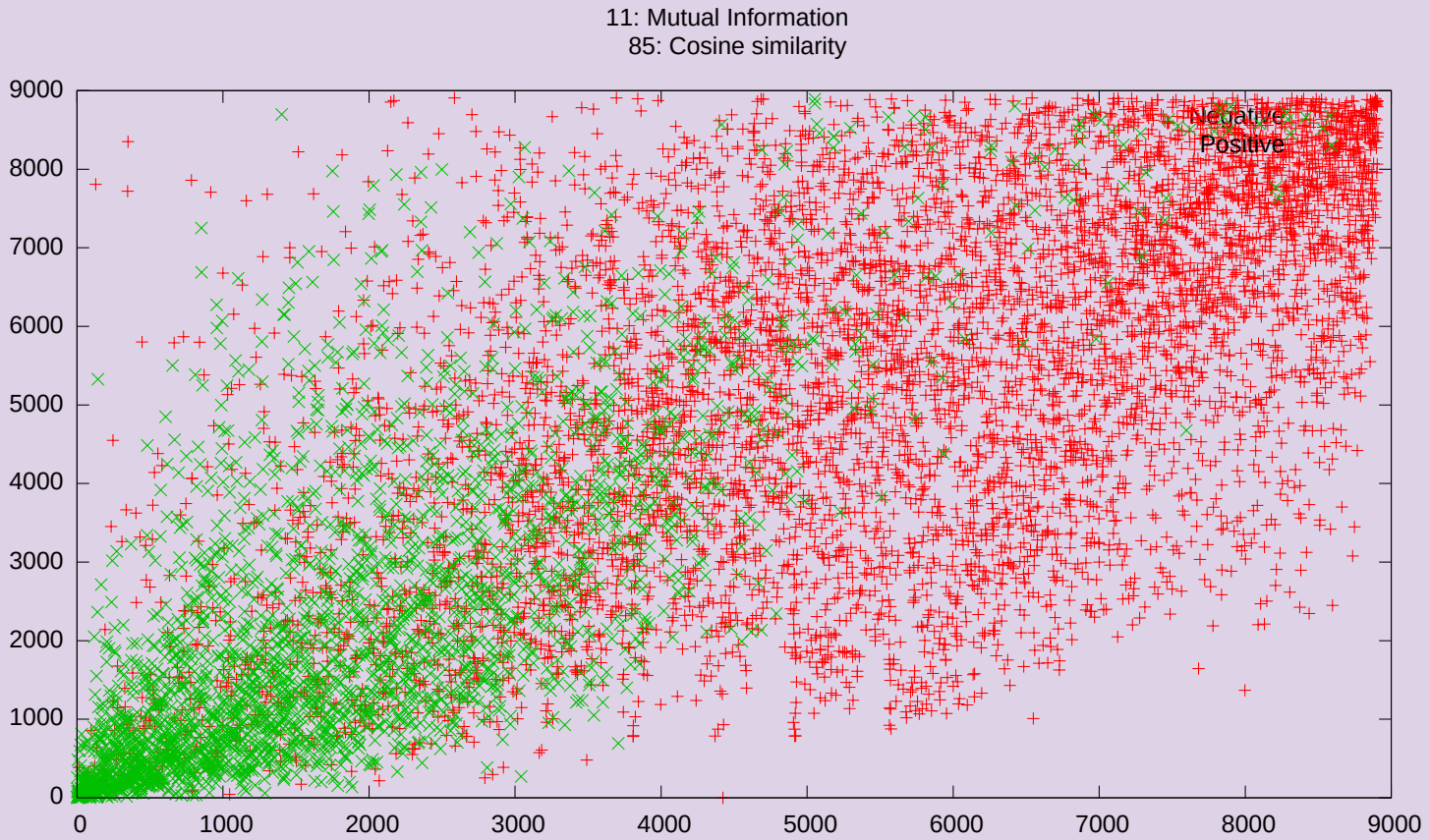
Evaluation results

Overview



Advanced Methods: motivation

Example: Mutual Information vs. Cosine context similarity



Advanced Methods: idea

Statistical learning problem

- for each bigram we get set of features (categories, scores etc.)

$$\mathbf{x_i} = (x_1, x_2, \ldots x_{90})$$

- each bigram we want to classify as collocation or noncolloc.

$$f(\mathbf{x_i}) = y_i, y_i = 0, 1$$

- so we are looking for function that minimizes a risk functional

$$\min \sum_i Q(f(\mathbf{x_i}), y_i)$$

Advanced Methods: idea cont.

Statistical learning problem

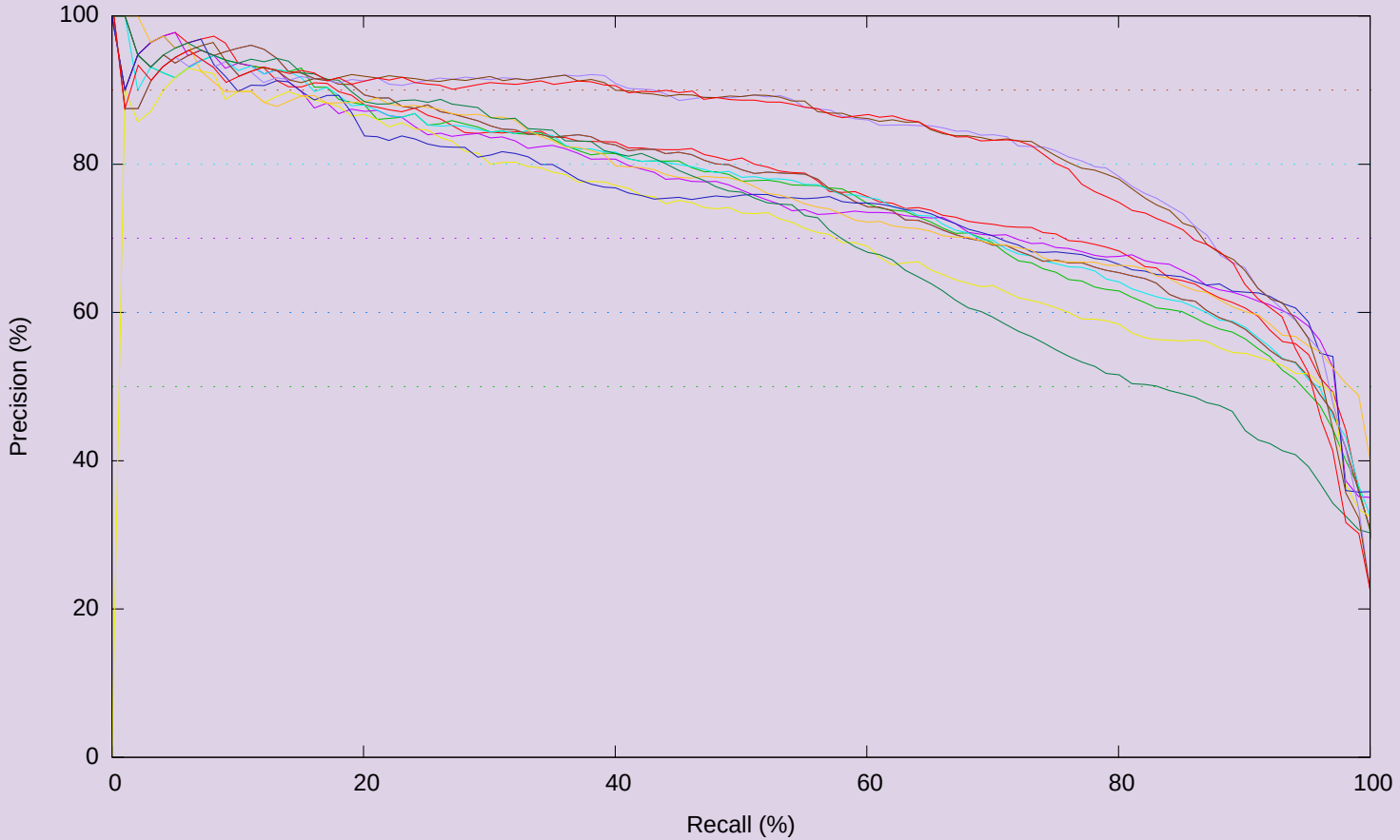
- but classification might be hard, what about regression?

$$f(\mathbf{x_i}) = y_y, y_i \in \langle 0, 1 \rangle$$

- Isn't R^{90} too much? What about feature selection?
- Yes! And how to do it?
  - Liner discriminant
  - General linear models - logistic regresion
  - Neural networks
  - Support vector Machines

Result

Support Vector Machines



Outline

Conclusion

Achived 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
- extremelly powerfull
- GNU GPL license
- www.r-project.org

Torch

- machine learning library
- C++, BSD license
- www.torch.ch