NPFL147

Statistical Methods in Natural Language Processing

5. Statistical properties of words

Pavel Pecina, Jindřich Helcl

11 November, 2025

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Course Segments

1. Introduction, probability, essential information theory
2. Statistical language modelling (n-gram)
3. Statistical properties of words
4. Word embeddings
5. Hidden Markov models, Tagging

Recap from Last Week

n-gram Language Model

- $(n\text{-}1)^{th}$ order Markov approximation → *n-gram Language Model*:

$$p(W) =^{df} \prod_{i=1..d} p(w_i | \overbrace{w_{i-n+1}, w_{i-n+2}, \ldots, w_{i-1}}^{\text{history}})$$

(prediction: w_i)

- In particular (assume vocabulary $|V| = 60k$):
  - 0-gram LM: uniform model, $p(w) = 1/|V|$, 1 parameter
  - 1-gram LM: unigram model, $p(w)$, 6×10^4 parameters
  - 2-gram LM: bigram model, $p(w_i|w_{i-1})$, 3.6×10^9 parameters
  - 3-gram LM: trigram model, $p(w_i|w_{i-2}, w_{i-1})$, 2.16×10^{14} parameters

Eliminating the Zero Probabilities: Smoothing

- Get new $p'(w)$ (same Ω): almost $p(w)$ but no zeros
- Discount w for (some) $p(w) > 0$: new $p'(w) < p(w)$

$$\Sigma_{w \in \text{discounted}} (p(w) - p'(w)) = D$$

- Distribute D to all w; $p(w) = 0$: new $p'(w) > p(w)$
  - possibly also to other w with low $p(w)$
- For some w (possibly): $p'(w) = p(w)$
- Make sure $\Sigma_{w \in \Omega} p'(w) = 1$
- There are many ways of ***smoothing***

Smoothing by Adding 1 and less than 1

- Simplest but not really usable:
  - Predicting words w from a vocabulary V, training data T:

    $\mathbf{p'(w|h) = (c(h,w) + 1) \ / \ (c(h) + |V|)}$
    - for non-conditional distributions: $p'(w) = (c(w) + 1) / (|T| + |V|)$
  - Problem if $|V| > c(h)$ (as is often the case; even $>> c(h)$!)
- Equally simple:
  - Predicting words w from a vocabulary V, training data T:

    $\mathbf{p'(w|h) = (c(h,w) + \lambda) \ / \ (c(h) + \lambda|V|), \lambda < 1}$
    - for non-conditional distributions: $p'(w) = (c(w) + \lambda) / (|T| + \lambda|V|)$

Good-Turing Smoothing

- Suitable for estimation from large data
  - Estimate probability of things that occur c times with the probability of things that occur $c+1$ times:

    $\mathbf{c' = (c+1) \times N(c+1)/N(c)}$
  - Full formula:

    $\mathbf{p_r(w) = (c(w) + 1) \times N(c(w) + 1) / (|T| \times N(c(w)))},$

    where $N(c)$ is the count of words with count c (count-of-counts)

    specifically, for $c(w) = 0$ (unseen words), $p_r(w) = N(1) / (|T| \times N(0))$
- good for small counts (< 5-10, where $N(c)$ is high)
- variants (see M&S)
- normalization! (so that we have $\Sigma_w\, p'(w) = 1$)

Language Model Interpolation

- Weight in less detailed distributions using $\lambda = (\lambda_0,\lambda_1,\lambda_2,\lambda_3)$:

$$p'_\lambda(w_i| w_{i-2},w_{i-1}) = \lambda_3 p_3(w_i| w_{i-2},w_{i-1}) + \lambda_2 p_2(w_i| w_{i-1}) + \lambda_1 p_1(w_i) + \lambda_0 /|V|$$

- Normalize:

$$\lambda_i > 0,\ \Sigma_{i=0..n} \lambda_i = 1 \qquad (\lambda_0 = 1 - \Sigma_{i=1..n} \lambda_i)\ (n=3)$$

- Estimation using MLE:
  - Fix the p_3, p_2, p_1 and $|V|$ parameters as estimated from training data
  - Find such $\{\lambda_i\}$ which minimizes the cross entropy (maximizes probability of data): $-(1/|D|)\Sigma_{i=1..|D|}\log_2(p'_\lambda(w_i|h_i))$

The Formulas

- Repeat: minimizing $-(1/|H|)\Sigma_{i=1..|H|}\log_2(p'_\lambda(w_i|h_i))$ over λ

$$p'_\lambda(w_i|h_i) = p'_\lambda(w_i|w_{i-2},w_{i-1}) =$$
$$= \lambda_3 p_3(w_i|w_{i-2},w_{i-1}) + \lambda_2 p_2(w_i|w_{i-1}) + \lambda_1 p_1(w_i) + \lambda_0/|V|$$

- "Expected Counts (of lambdas)": $j = 0,..,3$

$$c(\lambda_j) = \Sigma_{i=1..|H|}\ (\lambda_j p_j(w_i|h_i) / p'_\lambda(w_i|h_i))$$

- "Next λ": $j = 0,..,3$

$$\lambda_{j,next} = c(\lambda_j) / \Sigma_{k=0..3}\ (c(\lambda_k))$$

The (Smoothing) EM Algorithm

1. Start with some λ, such that $\lambda_j > 0$ for all $j \in 0,...,3$.
2. Compute "Expected Counts" for each λ_j.
3. Compute new set of λ_j, using the "Next λ" formula.
4. Start over at step 2, unless a termination condition is met.

- Termination condition: convergence of λ
  - Simply set an ε, and finish if $|\lambda_{j,old} - \lambda_{j,new}| < \varepsilon$ for each j (step 3).
- Guaranteed to converge:
  - Follows from Jensen's inequality, plus a technical proof

Statistical Properties of Words

Distributional Hypothesis

- Zellig Harris (1954) :
  - "*Words that occur in similar contexts tend to have similar meanings*"
  - i.e.. the meaning of a word can be inferred from the contexts in which it appears
- Context:
  - surrounding words, syntactic roles, co-occurrence patterns

Distributional Hypothesis

- Zellig Harris (1954) :
  - *"Words that occur in similar contexts tend to have similar meanings"*
  - i.e.. the meaning of a word can be inferred from the contexts in which it appears
- Context:
  - surrounding words, syntactic roles, co-occurrence patterns
- Examples:

Word	Common Contexts	Similar Words
cat	pet, fur, sleep, animal	dog, kitten, rabbit
bank	money, loan, account, interest	credit union, lender
river bank	water, flow, bridge, shore	coast, riverside

Distributional Hypothesis

- Zellig Harris (1954) :
  - *"Words that occur in similar contexts tend to have similar meanings"*
  - i.e.. the meaning of a word can be inferred from the contexts in which it appears
- Context:
  - surrounding words, syntactic roles, co-occurrence patterns
- Examples:

Word	Common Contexts	Similar Words
cat	pet, fur, sleep, animal	dog, kitten, rabbit
bank	money, loan, account, interest	credit union, lender
river bank	water, flow, bridge, shore	coast, riverside

- Implications:
  - Supports computational models that learn meaning from data.
  - Basis of distributional semantics (vector representations of words, embeddings)
  - Meaning emerges from usage patterns, not just dictionary definitions.

Collocations

- J. R. Firth (1957):
  - "*You shall know a word by the company it keeps.*"
  - "*Collocations of a given word are statements of the habitual or customary places of that word.*"
- M&S (Chapter 5):
  - *"A collocation is an expression consisting of two or more words that correspond to some conventional way of saying things."*
- Examples:
  - *strong tea, weapons of mass destruction*
  - *to make up, the rich and powerful*
- Valid or invalid?
  - *a stiff breeze,* but not *a stiff wind*
  - *broad daylight,* but not *bright daylight*

Properties of Collocations

- Typical properties/criterions of collocations:
  - non-compositionality
  - non-substitutability
  - non-modifiability

Properties of Collocations

- Typical properties/criterions of collocations:
  - non-compositionality
  - non-substitutability
  - non-modifiability
- Collocations usually cannot be translated word-by-word
  - e.g. *take a shower* → *osprchovat se,* but not *vzít sprchu*
- A phrase can be a collocation even if it is not consecutive
  - e.g. *knock … door*

Non-compositionality

- A phrase is compositional if the meaning can be predicted from the meaning of the parts.
  - e.g. *new companies*
- A phrase is non-compositional if the meaning cannot be predicted from the meaning of the parts
  - e.g. *hot dog*
- Collocations are not necessarily fully compositional in that there is usually an element of meaning added to the combination.
  - eg. *strong tea*
- Idioms are the most extreme examples of non-compositionality.
  - e.g. *to hear it through the grapevine*.

Non-substitutability, Non-modifiability

- (Near)-synonyms cannot substitute components of a collocation.
  - e.g. *yellow wine* vs. *white wine* even though it is kind of a yellowish white

- Many collocations cannot be freely modified with additional lexical material or through grammatical transformations
  - e.g. *white wine*, but not *whiter wine*
  - *mother in law*, but not *mother in laws*

Collocations of Special Interest

- Idioms: really fixed phrases
  - *kick the bucket, birds-of-a-feather, run for office*
- Proper names: difficult to recognize even with lists
  - *Tuesday (person's name), May, Winston Churchill, IBM, Inc.*
- Numerical expressions
  - *Monday Oct 04 1999, two thousand seven hundred fifty*
- Phrasal verbs
  - Separable parts:
  - *look up, take off*

Word Association and Co-occurence

- Does not fall under “collocation”
- Interesting just because it does often [rarely] appear together or in the same (or similar) context.
- Examples:
  - *(doctors, nurses)*
  - *(hardware, software)*
  - *(gas, fuel)*
  - *(hammer, nail)*
  - *(communism, free speech)*

Further Notions

- Synonymy: different form/word, same meaning:
  - *notebook / laptop*
- Antonymy: opposite meaning:
  - *new/old, black/white, start/stop*
- Homonymy: same form/word, different meaning:
  - "true" (random, unrelated): *can (aux. verb / can of Coke)*
  - related: polysemy; *notebook, shift, grade, ...*
- Other:
  - Hyperonymy/Hyponymy: general vs. special: *vehicle/car*
  - Meronymy/Holonymy: whole vs. part: *body/leg*

How to Find Collocations?

- Frequency (simplest method)
  - plain
  - filtered
- Mean and variance of the distance between focal word and collocating word
- Hypothesis testing
  - *t* test
  - $\chi 2$ test
- Pointwise Mutual Information

Frequency

- Count n-grams; high frequency n-grams are candidates:

Frequency

- Count n-grams; high frequency n-grams are candidates:

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Frequency

- Count n-grams; high frequency n-grams are candidates:
  - mostly function words
    - *of the*
  - frequent names
    - *New York*

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Frequency

- Count n-grams; high frequency n-grams are candidates:
  - mostly function words
    - *of the*
  - frequent names
    - *New York*
- Filtering:
  - Stop list: words/forms which (we think) cannot be a part of a collocation
    - *a, the, and, or, but, not, …*
  - Part of Speech (possible collocation patterns)
    - A+N, N+N, N+of+N, …

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

Frequency

- Count n-grams; high frequency n-grams are candidates:
  - mostly function words
    - *of the*
  - frequent names
    - *New York*
- Filtering:
  - Stop list: words/forms which (we think) cannot be a part of a collocation
    - *a, the, and, or, but, not, …*
  - Part of Speech (possible collocation patterns)
    - A+N, N+N, N+of+N, …

$C(w^1\ w^2)$	w^1	w^2	tag pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

POS Tag Patterns for Collocation Filtering

A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

- Defined by Justeson and Katz (1995)

Collocational Window

- Many collocations occur at variable distances.
- A collocational window needs to be set to locate these.
    - e.g. by maximum offset between the components
- Frequency based approach can't be used.
- Examples:
    - *she knocked on his door*
    - *they knocked at the door*
    - *100 women knocked on the Donaldson's door*
    - *a man knocked on the metal front door*

Mean and Variance

- The mean μ is the average offset (signed distance) between two words in the corpus.
- The variance σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

  - n is the number of times the two words co-occur
  - d_i is the offset for co-occurrence i
  - μ is the mean, σ is the standard deviation

Mean and Variance

- The mean μ is the average offset (signed distance) between two words in the corpus.
- The variance σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

  - n is the number of times the two words co-occur
  - d_i is the offset for co-occurrence i
  - μ is the mean, σ is the standard deviation

- Mean and variance characterize the distribution of distances between two words in a corpus.
  - High variance means that co-occurrence is mostly by chance
  - Low variance means that the two words usually occur at about the same distance.

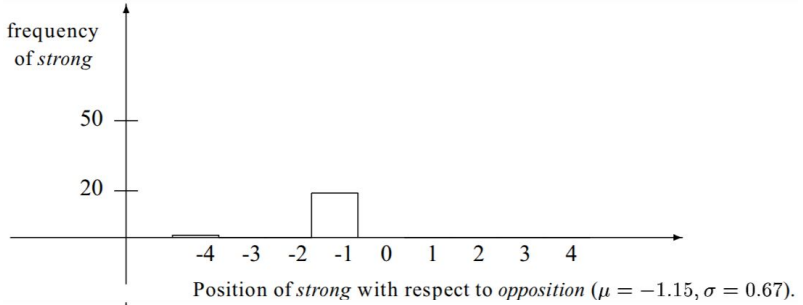
Mean and Variance

- The mean μ is the average offset (signed distance) between two words in the corpus.
- The variance σ^2

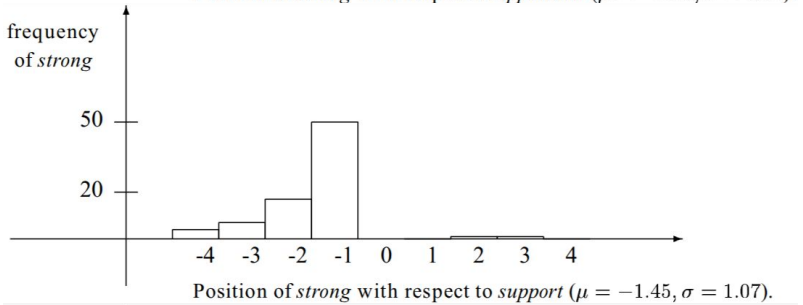
$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

  - n is the number of times the two words co-occur
  - d_i is the offset for co-occurrence i
  - μ is the mean, σ is the standard deviation

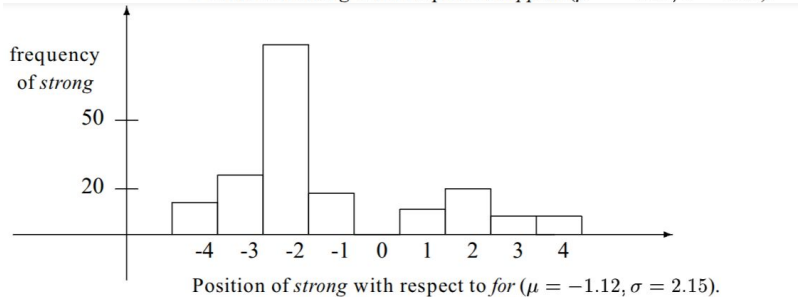
- Mean and variance characterize the distribution of distances between two words in a corpus.
  - High variance means that co-occurrence is mostly by chance
  - Low variance means that the two words usually occur at about the same distance.



Position of *strong* with respect to *opposition* ($\mu = -1.15, \sigma = 0.67$).



Position of *strong* with respect to *support* ($\mu = -1.45, \sigma = 1.07$).



Position of *strong* with respect to *for* ($\mu = -1.12, \sigma = 2.15$).

Mean and Variance: An example

- For the *knock* - *door* example sentences the sample mean is:

$$\mu = \frac{1}{4}(3 + 3 + 5 + 5) = 4.0$$

- And the standard deviation:

$$\sigma = \sqrt{\frac{1}{3}\left((3 - 4.0)^2 + (3 - 4.0)^2 + (5 - 4.0)^2 + (5 - 4.0)^2\right)} \approx 1.15$$

Mean and Variance: An example cont'd

σ	μ	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

Hypothesis Testing

- Two words can co-occur by chance
  - High frequency and low variance can be accidental
- Hypothesis Testing measures the confidence that this co-occurrence was really due to association, and not just due to chance.

Hypothesis Testing

- Two words can co-occur by chance
  - High frequency and low variance can be accidental
- Hypothesis Testing measures the confidence that this co-occurrence was really due to association, and not just due to chance.
- Formulate a null hypothesis H_0 that there is no association between the words beyond chance occurrences:
  - H_0 states what should be true if two words do not form a collocation.
  - If H_0 can be rejected, the words do not co-occur by chance, and they form a collocation
- Compute the probability p that the event would occur if H_0 were true:
  - reject H_0 if p is too low (typically beneath a significance level of $p < 0.05, 0.01, \ldots$)
  - retain H_0 as possible otherwise.

t-test (Student's *t*-test)

- The test looks at the difference between the observed and expected means, scaled by the variance of the data, and tells us how likely one is to get a sample of that mean and variance, assuming that the sample is drawn from a normal distribution with mean μ.

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}}$$

  - Where $\bar{x}$ is the real data mean (observed in data)
  - s^2 is the variance
  - N is the sample size
  - μ is the mean of the distribution (expected under H_0)

Finding Collocations by *t*-test

- Think of the text corpus as a long sequence of N bigrams, and the samples are then indicator random variables with:
  - value 1 when the bigram of interest occur ($x_i = 1$ if $(w_{i-1}, w_i) =$ *"New York"*)
  - 0 otherwise.
- Example:

  W = *New York residents often say New York never sleeps .*

  $X = 1, 0, 0, 0, 0, 1, 0, 0$
- The t-test and other statistical tests are useful as methods for ranking collocations.
  1. Determine the expected mean
  2. Measure the observed mean
  3. Run the t-test

t-test: An example

- In our corpus:
  - *new* occurs 15,828 times, *companies* 4,675 times, 14,307,668 tokens overall.
  - *new companies* occurs 8 times among the 14,307,667 bigrams
- H_0: "*new companies*" occur at random:
  - $P(\textit{new companies}) = P(\textit{new})P(\textit{companies}) = \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$
- $\mu = 3.615 \text{ x } 10^{-7}$, $s^2 = p(1\text{-}p) \approx p$, $\bar{x} = \frac{8}{14307668} \approx 5.591 \times 10^{-7}$

$$t = \frac{\bar{x} - \mu}{\sqrt{\frac{s^2}{N}}} \approx \frac{5.59110^{-7} - 3.61510^{-7}}{\sqrt{\frac{5.59110^{-7}}{14307668}}} \approx 0.999932$$

- t value of 0.999932 is not larger than 2.576, the critical value for $a{=}0.005$.
- We cannot reject H_0 (*new* and *companies* occur independently) and do not form a collocation.

t-test: An example

t	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
4.4720	30	117	20	Agatha	Christie
4.4720	77	59	20	videocassette	recorder
4.4720	24	320	20	unsalted	butter
2.3714	14907	9017	20	first	made
2.2446	13484	10570	20	over	many
1.3685	14734	13478	20	into	them
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

- The *t*-test applied to 10 bigrams that occur with frequency 20.

Pearson's χ^2 (chi-square) test

- *t*-test assumes that probabilities are approximately normally distributed, which is not true in general
- The χ^2 test doesn't make this assumption
- The χ^2 test compares the observed frequencies with the frequencies expected for independence
  - if the difference between observed and expected frequencies is large, then we can reject the null hypothesis of independence
- The χ^2 test relies on co-occurrence table and computes:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2 test: An example

- Observed occurrences:

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

- The χ^2 statistic sums differences between observed and expected values in all cells of the table, scaled by the magnitude of the expected values:

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

  - where i ranges over rows of the table, j ranges over columns, O_{ij} is the observed value for cell (i, j) and E_{ij} is the expected value.

χ^2 test: An example

- Observed values O:
  - e.g. $O_{1,1} = 8$

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

- Expected values E are determined from marginal probabilities:
  - e.g. $E_{1,1}$ = expected frequency for *new companies*, determined by:

$$E_{1,1} = P(new,*)\ P(*,companies) = \frac{8+4667}{N} \times \frac{8+15820}{N} \times N \approx 5.2$$

- χ^2 is then determined as 1.55
  - $\chi^2 = 3.8$ for probability level of $a = 0.05$ (look up in significance table)
  - $1.55 < 3.8$ we cannot reject null hypothesis *new companies* is not a collocation.

Pointwise Mutual Information

- An information-theoretically motivated measure for discovering interesting collocations is pointwise mutual information
- It is a measure of how much one word tells us about the other:

$$
\begin{aligned}
I(x', y') &= \log_2 \frac{P(x'y')}{P(x')P(y')} \\
&= \log_2 \frac{P(x'|y')}{P(x')} \\
&= \log_2 \frac{P(y'|x')}{P(y')}
\end{aligned}
$$

Pointwise Mutual Information

- An information-theoretically motivated measure for discovering interesting collocations is pointwise mutual information
- It is a measure of how much one word tells us about the other:

$$I(x', y') = \log_2 \frac{P(x'y')}{P(x')P(y')}$$

"observed"

"expected (under H_0)"

$$= \log_2 \frac{P(x'|y')}{P(x')}$$

$$= \log_2 \frac{P(y'|x')}{P(y')}$$

- This is NOT the MI as defined in Information Theory
  - MI average of the PMIs (for all the values)

PMI: An example

$I(w^1, w^2)$	$C(w^1)$	$C(w^2)$	$C(w^1\ w^2)$	w^1	w^2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
16.31	30	117	20	Agatha	Christie
15.94	77	59	20	videocassette	recorder
15.19	24	320	20	unsalted	butter
1.09	14907	9017	20	first	made
1.01	13484	10570	20	over	many
0.53	14734	13478	20	into	them
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

- 10 bigrams that occur with frequency 20, ranked according to PMI.

Moodle Quiz

https://dl1.cuni.cz/course/view.php?id=18547