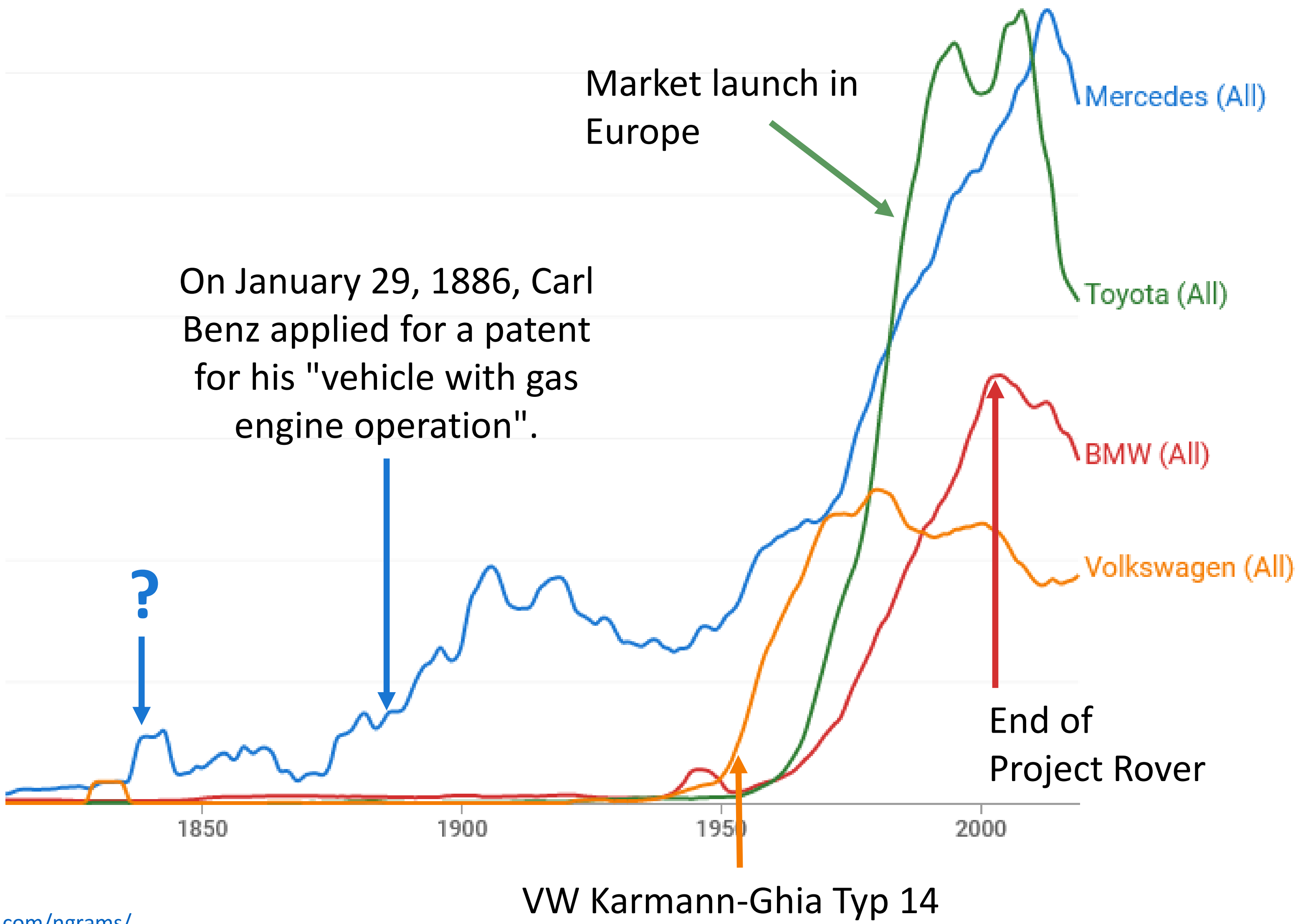
Word frequencies



http://books.google.com/ngrams/

Technische Hochschule Ingolstadt

How many bits do you need?

Information content IC of a word, sentence, text w:

$$IC(w) = log_2 \frac{1}{P(w)} = -log_2 P(w)$$

How many bits do you need?

Information content IC of a word, sentence, text w:

$$IC(w) = log_2 \frac{1}{\boxed{P(w)}} = -log_2 P(w)$$

Probability of occurrence or how likely is w. Typically calculated with a language model P

How many bits do you need?

Information content IC of a word, sentence, text w:

$$IC(w) = log_2 \frac{1}{P(w)} = -log_2 P(w)$$

Example:

P(w) = ½ => IC(w) = ?

Technische Hochschule Ingolstadt

How many bits do you need?

Information content IC of a word, sentence, text w:

$$IC(w) = log_2 \frac{1}{P(w)} = -log_2 P(w)$$

Example:

P(w) = ½ => IC(w) = ?

P(v) = 1/16 => IC(v) = ?

Technische Hochschule Ingolstadt

Entropy

„Measure of the average information content of a message, i.e. the expected value of the information content:“

$$H(X) = -\sum_{w \in X} p(w) \log_2 p(w)$$

Information content of w in X

Claude E. Shannon in about 1948

Entropy

„Measure of the average information content of a message, i.e. the expected value of the information content:“

$$H(X) = -\sum_{w \in X} p(w) \log_2 p(w)$$

Information content of w in X

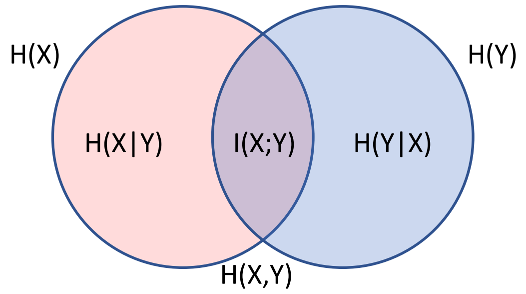
How many letters does it take anyway?

Claude E. Shannon in about 1948

Conditional Entropy

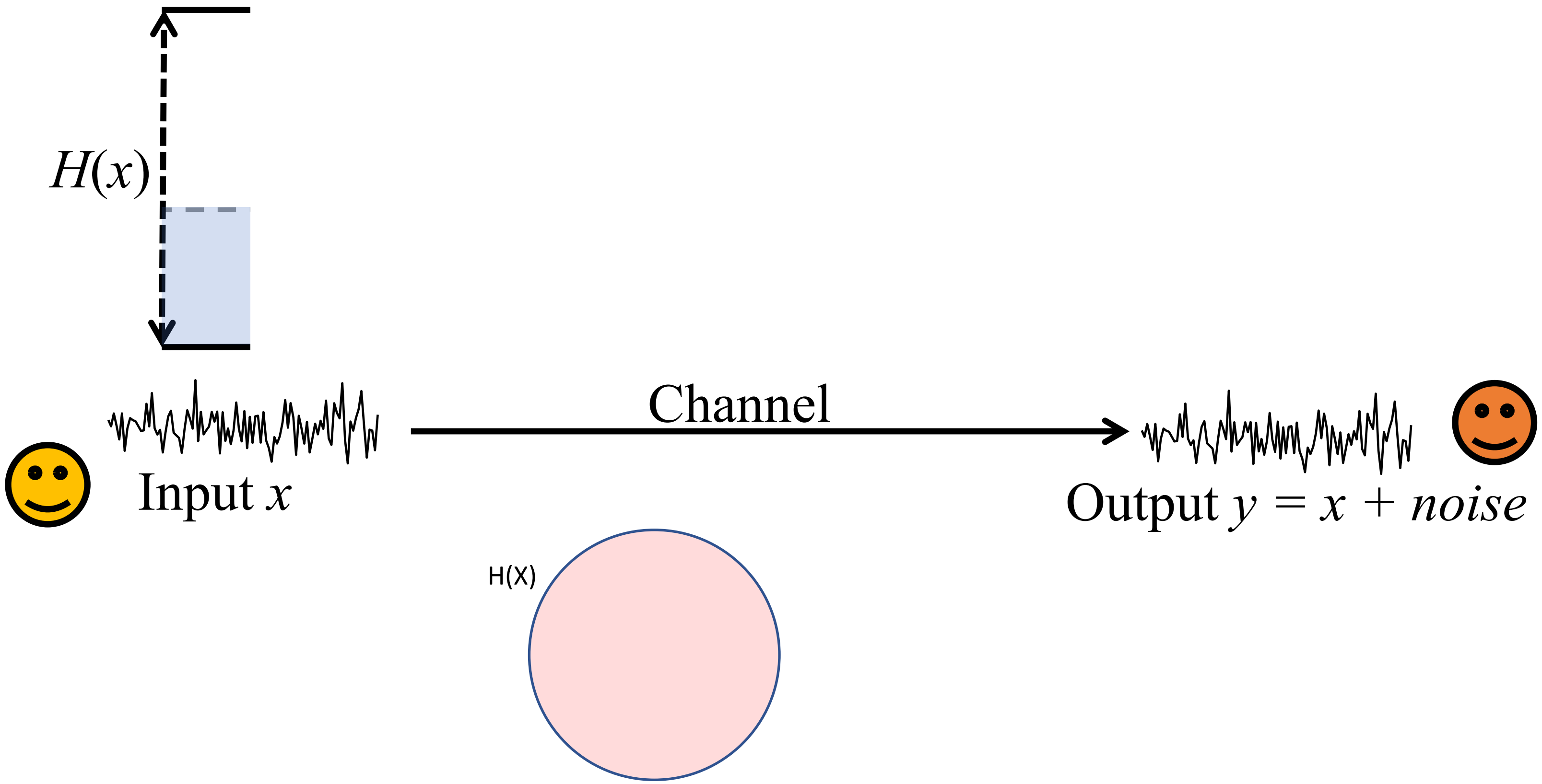
"Measure of the 'uncertainty' about the value of a random variable Y that remains after the outcome of another random variable X becomes known. "

$$
\begin{aligned}
H(Y|X) &= \sum_{w \in X} p(w) H(Y|X=w) \\
&= \sum_{w \in X} p(w) \Big[-\sum_{v \in Y} p(v|w) \log_2 p(v|w) \Big]
\end{aligned}
$$



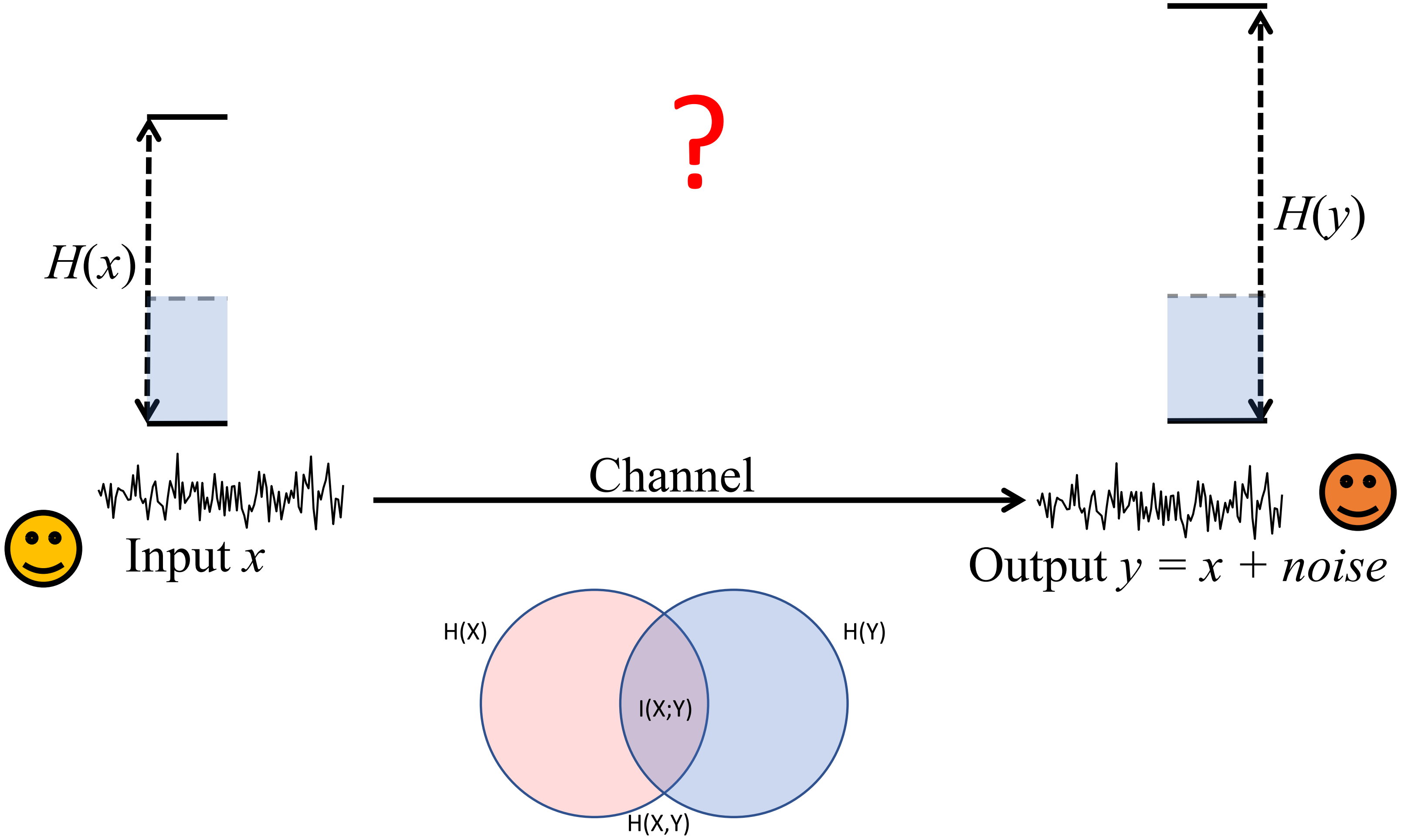
https://de.wikipedia.org/wiki/Bedingte_Entropie

Conditional Entropy

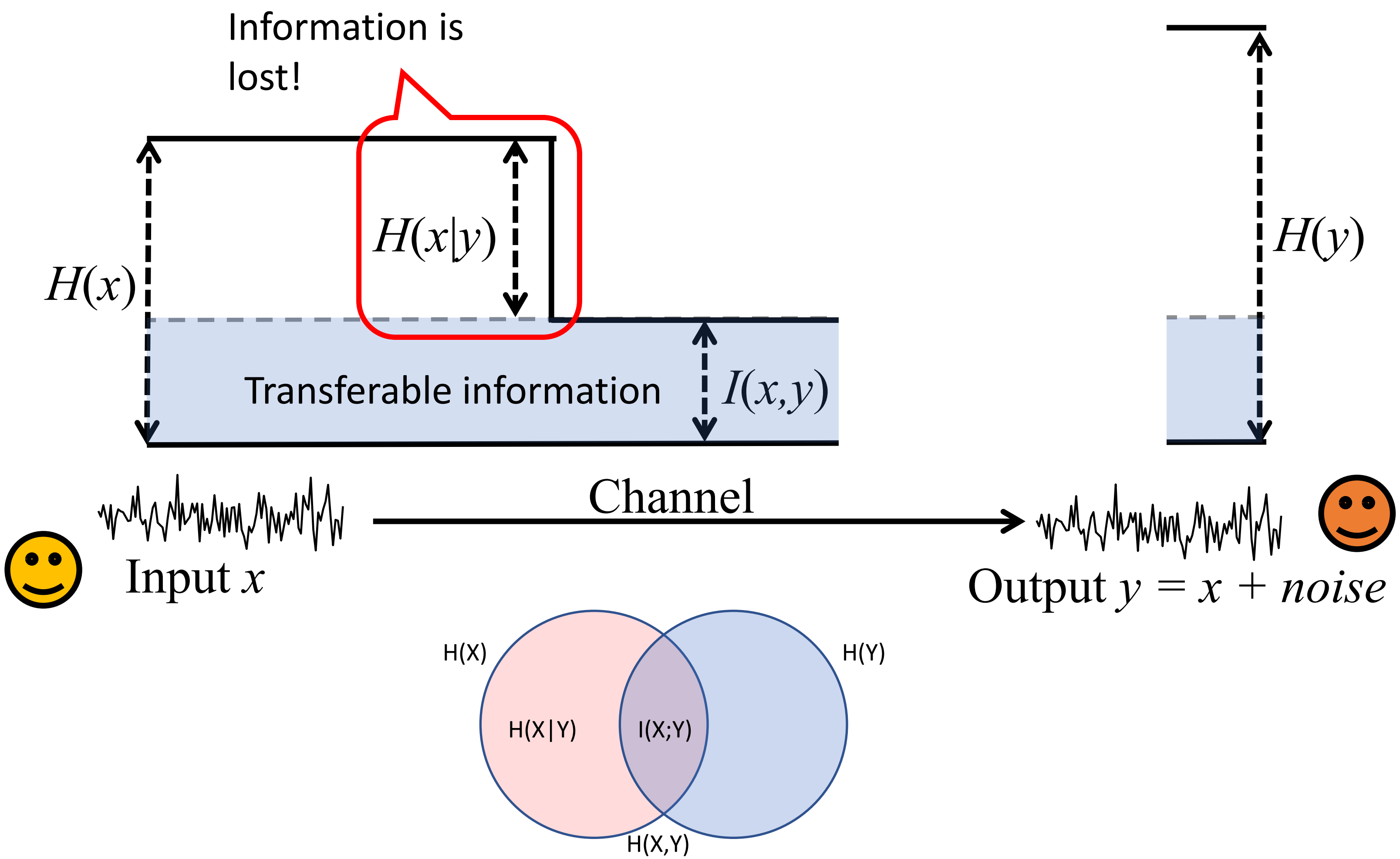


Technische Hochschule Ingolstadt

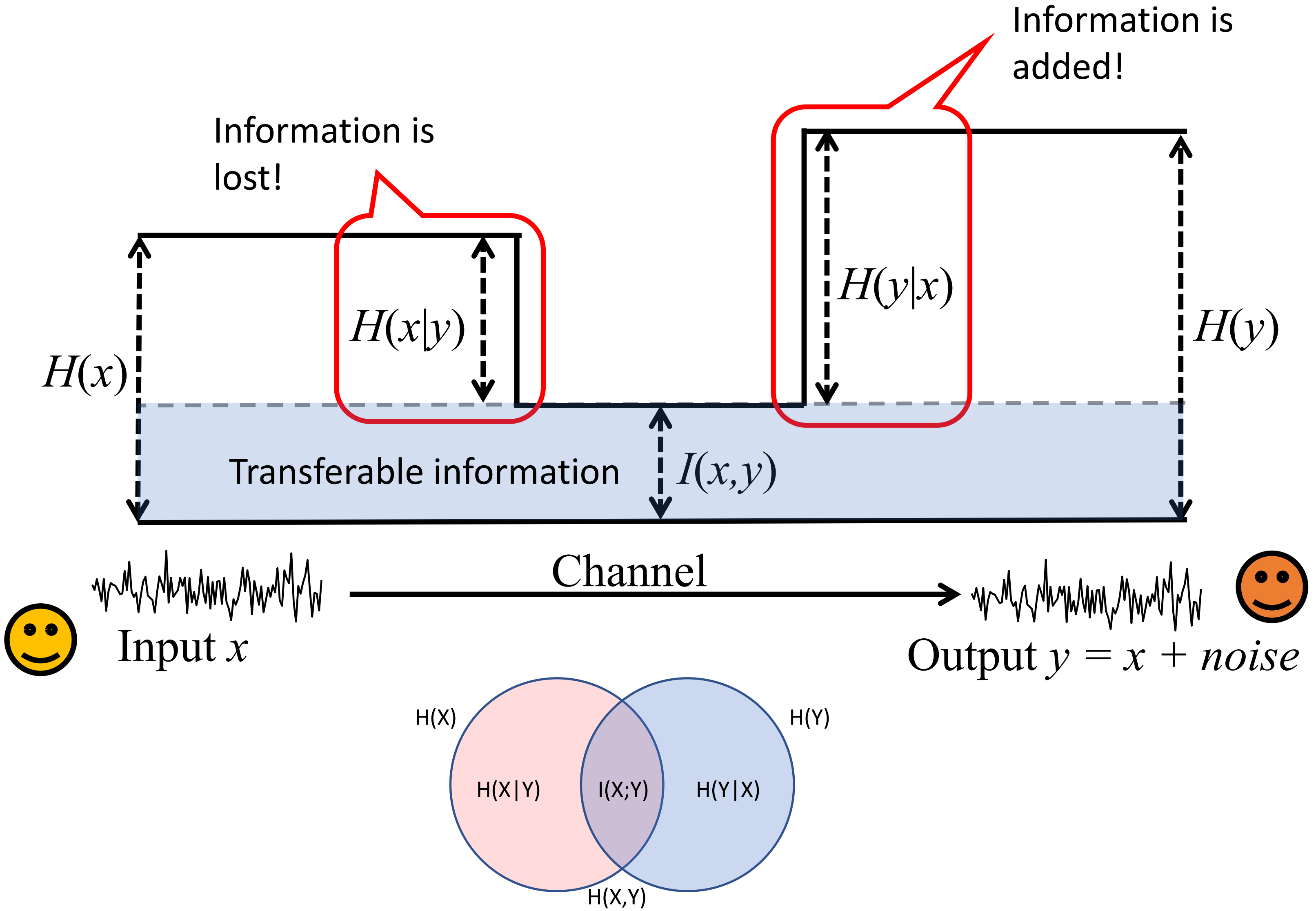
Conditional Entropy



Conditional Entropy



Conditional Entropy



Technische Hochschule Ingolstadt

Conditional Entropy

"Measure of the 'uncertainty' about the value of a random variable Y that remains after the outcome of another random variable X becomes known. "

$$\begin{aligned} H(Y|X) &= \sum_{w \in X} p(w) H(Y|X = w) \\ &= \sum_{w \in X} p(w) [- \sum_{v \in Y} p(v|w) \log_2 p(v|w)] \end{aligned}$$

Comparison!

$$H(X) = - \sum_{w \in X} p(w) \log_2 p(w)$$

Technische Hochschule Ingolstadt

Conditional Entropy

"Measure of the 'uncertainty' about the value of a random variable Y that remains after the outcome of another random variable X becomes known. "

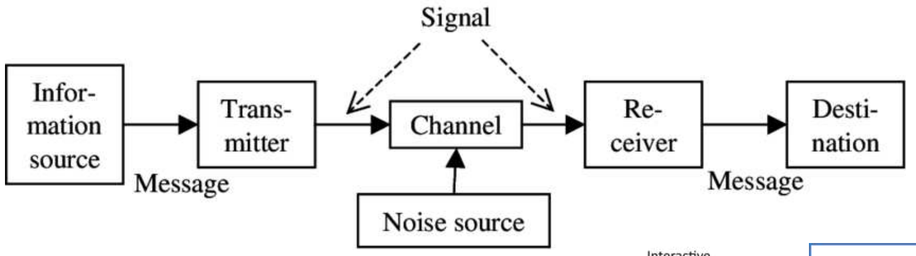
$$
\begin{aligned}
H(Y|X) &= \sum_{w \in X} p(w) H(Y|X = w) \\
&= \sum_{w \in X} p(w) [- \sum_{v \in Y} p(v|w) \log_2 p(v|w)]
\end{aligned}
$$

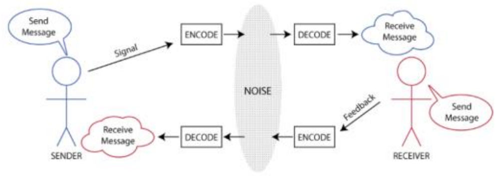
H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)

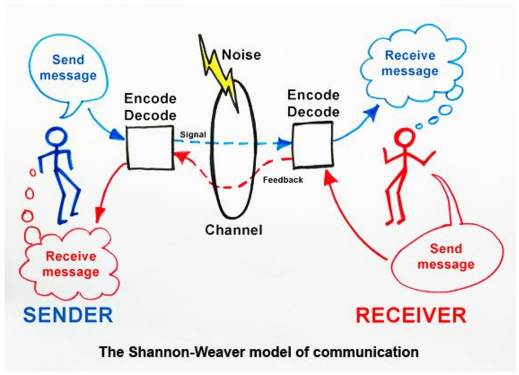
https://spinfo.phil-fak.uni-koeln.de/fileadmin/spinfo/ql/QL_04.pdf

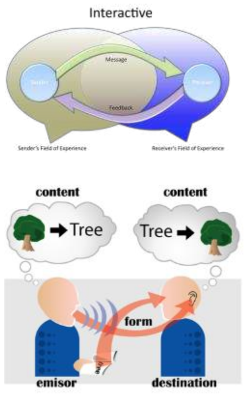
Shannon–Weaver model

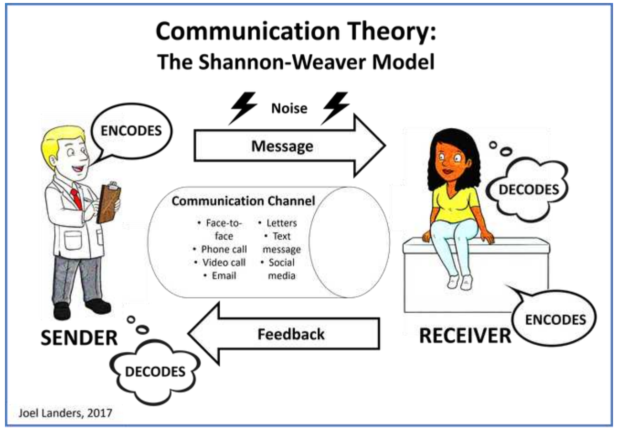
"mother of all models."*











*: *Erik Hollnagel and David D. Woods (2005). Joint Cognitive Systems: Foundations of Cognitive Systems Engineering. Boca Raton, FL: Taylor & Francis. ISBN 978-0-8493-2821-3.*

Relative Entropy

"Relative entropy gives the divergence of two probability distributions p and m for the same event space X."

$$D(p||m) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{m(x)}$$

Also called "Kullback-Leibler divergence".

Divergence := the difference
https://de.wikipedia.org/wiki/Kullback-Leibler-Divergenz

Relative Entropy

"Relative entropy gives the divergence of two probability distributions p and m for the same event space X."

$$D(p||m) = \sum_{x \in X} p(x) \log_2 \frac{p(x)}{m(x)}$$

Also called "Kullback-Leibler divergence".

Used e.g. for training neural networks

Divergenz := die Verschiedenheit
https://de.wikipedia.org/wiki/Kullback-Leibler-Divergenz

Technische Hochschule Ingolstadt

Conditional Relative Entropy

"Semantic similarities between pairs of words based on similar distribution in the training data can thus be detected."

$$D(p(y|x)\|m(y|x)) = \sum_x p(x) \sum_y p(y|x) \log_2 \frac{p(y|x)}{m(y|x)}$$

D(p(x,y) || q(x,y)) = D(p(x) || q(x)) + D(p(y|x) || q(y|x))

https://www.cs.princeton.edu/courses/archive/fall11/cos597D/L03.pdf

Cross Entropy

"The goal in evaluating language models is to determine the goodness of fit of a model to the given word sequence. However, the actual underlying distribution of the data is unknown."

Cross Entropy

"The goal in evaluating language models is to determine the goodness of fit of a model to the given word sequence. However, the actual underlying distribution of the data is unknown."

Relative entropy between the unknown distribution and our language model.

$$H(W_{1,n}, \boxed{m}) = H(W_{1,n}) + \boxed{D(p||m)}$$

Language model

Cross Entropy

"The goal in evaluating language models is to determine the goodness of fit of a model to the given word sequence. However, the actual underlying distribution of the data is unknown."

Relative entropy between the unknown distribution and our language model.

$$
\begin{aligned}
H(W_{1,n}, \boxed{m}) &= H(W_{1,n}) + \boxed{D(p||m)} \\
&= -\sum_{w_{1,n}} p(w_{1,n}) \log_2 m(w_{1,n})
\end{aligned}
$$

Language model

Cross entropy rate

"Normalization to sequence lengths yields the length-independent cross entropy rate."

$$
\begin{aligned}
H(W_{1,n}, m) &= H(W_{1,n}) + D(p||m) \\
&= -\sum_{w_{1,n}} p(w_{1,n}) \log_2 m(w_{1,n})
\end{aligned}
$$

$$\Downarrow$$

$$H_{\text{rate}}(W_{1,n}, m) = -\frac{1}{n} \sum_{w_{1,n}} p(w_{1,n}) \log_2 m(w_{1,n})$$

Cross entropy for a language L

"Cross entropy for an infinite word sequence."

$$H(L,m) = -\lim_{n\to\infty} \frac{1}{n} \sum_{w_{1,n}} p(w_{1,n}) \log_2 m(w_{1,n})$$

$$H_{\text{rate}}(W_{1,n},m) = -\frac{1}{n} \sum_{w_{1,n}} p(w_{1,n}) \log_2 m(w_{1,n})$$

Cross entropy for a language L

"The complete population of the language L is completely covered, i.e. all possible and impossible sentences. The weighted mean is therefore no longer necessary. "

$$H(L,m) = -\lim_{n\to\infty} \frac{1}{n} \log_2 m(w_{1,n})$$

It is assumed thatlanguage is an "ergodic stochastic process" ...

$$H(L,m) = -\lim_{n\to\infty} \frac{1}{n} \sum_{w_{1,n}} \cancel{p}(w_{1,n}) \log_2 m(w_{1,n})$$

$$H_{\text{rate}}(W_{1,n},m) = -\frac{1}{n} \sum_{w_{1,n}} p(w_{1,n}) \log_2 m(w_{1,n})$$

Cross entropy for a language L

"Since not all word sequences of L are available, the cross entropy can only be approximated."

$$H(L, m) = -\lim_{n \to \infty} \frac{1}{n} \log_2 m(w_{1,n})$$

?

Cross entropy for a language L

"Since not all word sequences of L are available, the cross entropy can only be approximated."

$$H(L,m) = -\cancel{\lim_{n\to\infty}} \frac{1}{n} \log_2 m(w_{1,n})$$

$$H(L,m) \approx -\frac{1}{n} \log_2 m(w_{1,n})$$

"A confident prediction regarding word sequence, is reflected in a lower cross entropy."

Technische Hochschule Ingolstadt

Perplexity

"Standard Evaluation for Language Models. The better the model, the lower the perplexity scores".

$$PP(L,m) = 2^{H(L,m)}$$

$$PP(L,m) = 2^{-\frac{1}{n}\log_2 m(w_{1,n})}$$

Perplexity as „branching factor“

10 digits: 0,1,2,3,4,5,6,7,8,9

$$P(d) = 2^{-\frac{1}{1}\log_2\left(\frac{1}{10}\right)^1}$$

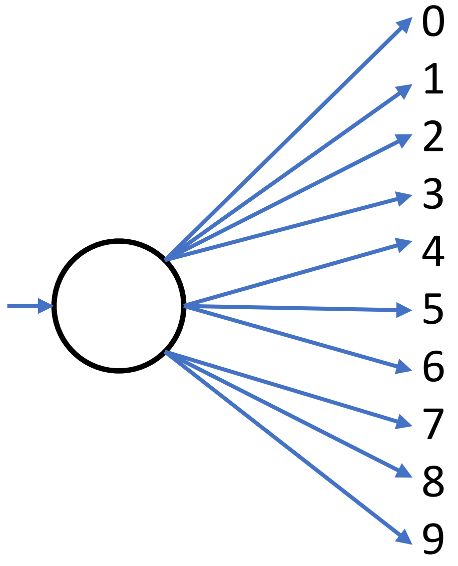
$$PP(L,m) = 2^{-\frac{1}{n}\log_2 m(w_{1,n})}$$

Perplexity as „branching factor“

10 digits: 0,1,2,3,4,5,6,7,8,9

$$P(d) = 2^{-\frac{1}{1}\log_2\left(\frac{1}{10}\right)^1}$$
$$= 10$$

branching factor, d.h. 10
Outgoing edges



Language model evaluation

Information theory was developed in the 1940s from the interest in maximizing the amount of information transmitted over a noisy channel (e.g., radio link). For this purpose, the compression capacity of the data (with the help of entropy) and the channel capacity have to be determined in general.

https://www.phonetik.uni-muenchen.de/~reichelu/kurse/stat_lm/script_reichel_stat_lm.pdf

Perplexity

„Standard Evaluierung für Sprachmodelle. Je besser das Modell, desto niedriger die Perplexitätswerte“

$$PP(L,m) = 2^{H(L,m)}$$

$$PP(L,m) = 2^{-\frac{1}{n}\log_2 m(w_{1,n})}$$

Whatever you do in

Language Modeling

it‘s all about

minimizing the perplexity.

Language models estimation

"For example, if in a text corpus the most frequent type occurs 10000 times, then already the 10000 most frequent type occurs only once*.So we have a Large number of rare events problem"

*sie Zipf's Law.

Technische Hochschule Ingolstadt

2-gram

P(<s> Diese Vorlesung ist spannend </s>) ≈

P(<s>)

·P(Diese|<s>)

·P(Vorlesung| Diese)

·P(ist| Vorlesung)

·P(spannend| ist)

·P(</s>| spannend)

$P(w_i|w_{i-1}) =$?

Remember:

$P(A \cap B) = P(A \mid B)\, P(B)$

2-gram

P(<s> Diese Vorlesung ist spannend </s>) ≈

P(<s>)

·P(Diese|<s>)

·P(Vorlesung| Diese)

·P(ist| Vorlesung)

·P(spannend| ist)

·P(</s>| spannend)

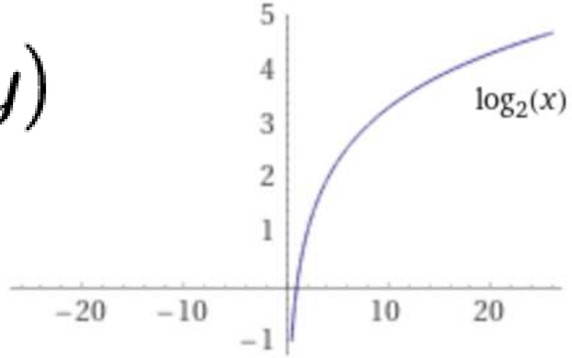
Maximum Likelihood Estimate (MLE)

|.| := Count occurrences of.

$$P(w_i|w_{i-1}) = \frac{P(w_i w_{i-1})}{P(w_{i-1})} = \frac{\frac{|w_i w_{i-1}|}{|corpus|}}{\frac{|w_{i-1}|}{|corpus|}} = \frac{|w_i w_{i-1}|}{|w_{i-1}|}$$

Comment: log-probabilities

$$\log_b(x \cdot y) = \log_b(x) + \log_b(y)$$



- Velocity:

  \+ is faster than *

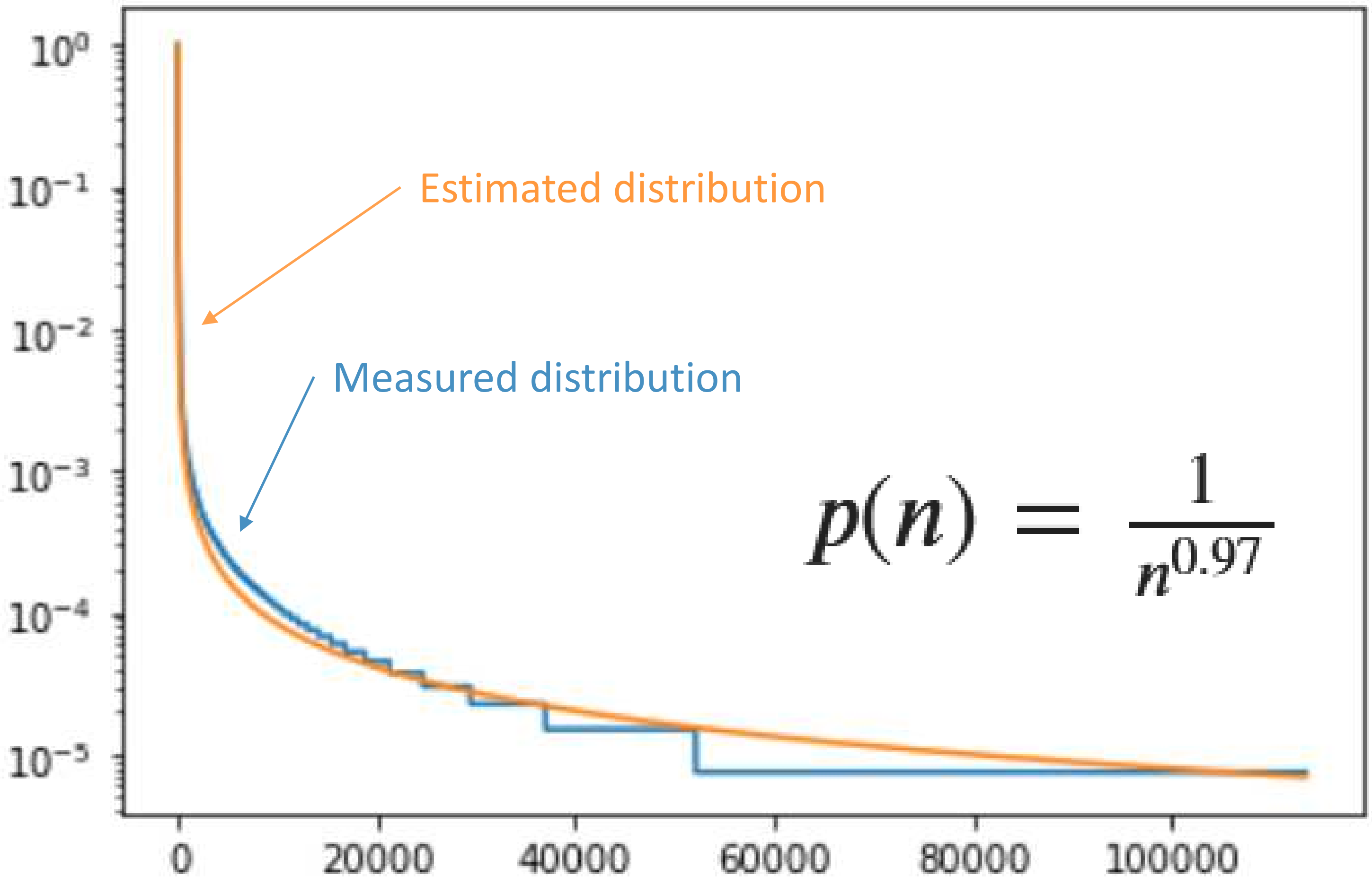
  log(.) only needs to be calculated once

- Accuracy: Numerical Stable on GPU/CPU

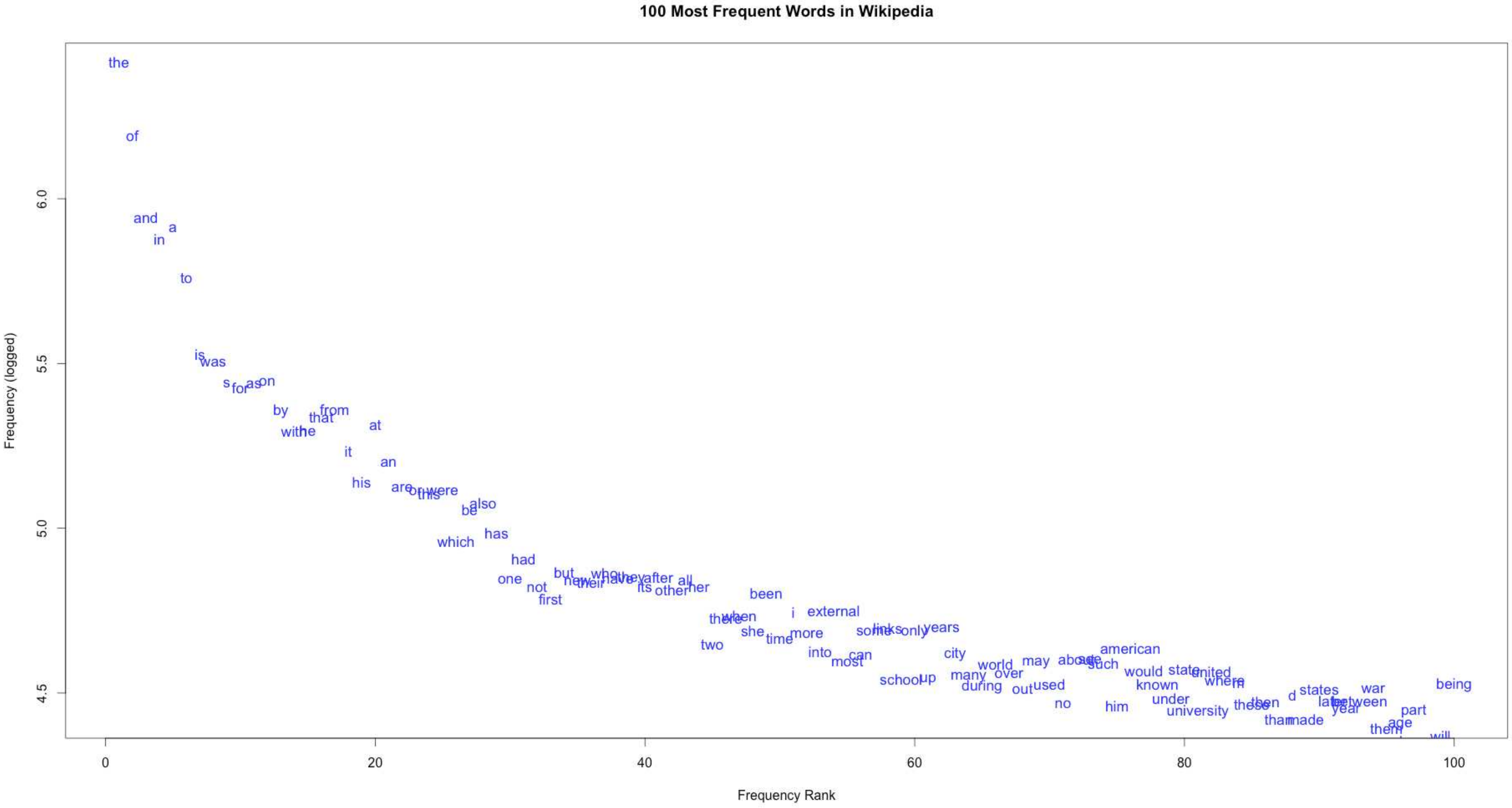
- *Many distributions are exponential in nature, i.e. you save the "e" to the power of...*

https://en.wikipedia.org/wiki/Log_probability
https://en.wikipedia.org/wiki/Floating-point_arithmetic
https://de.wikipedia.org/wiki/Stabilit%C3%A4t_(Numerik)

Large number of rare events-Problem

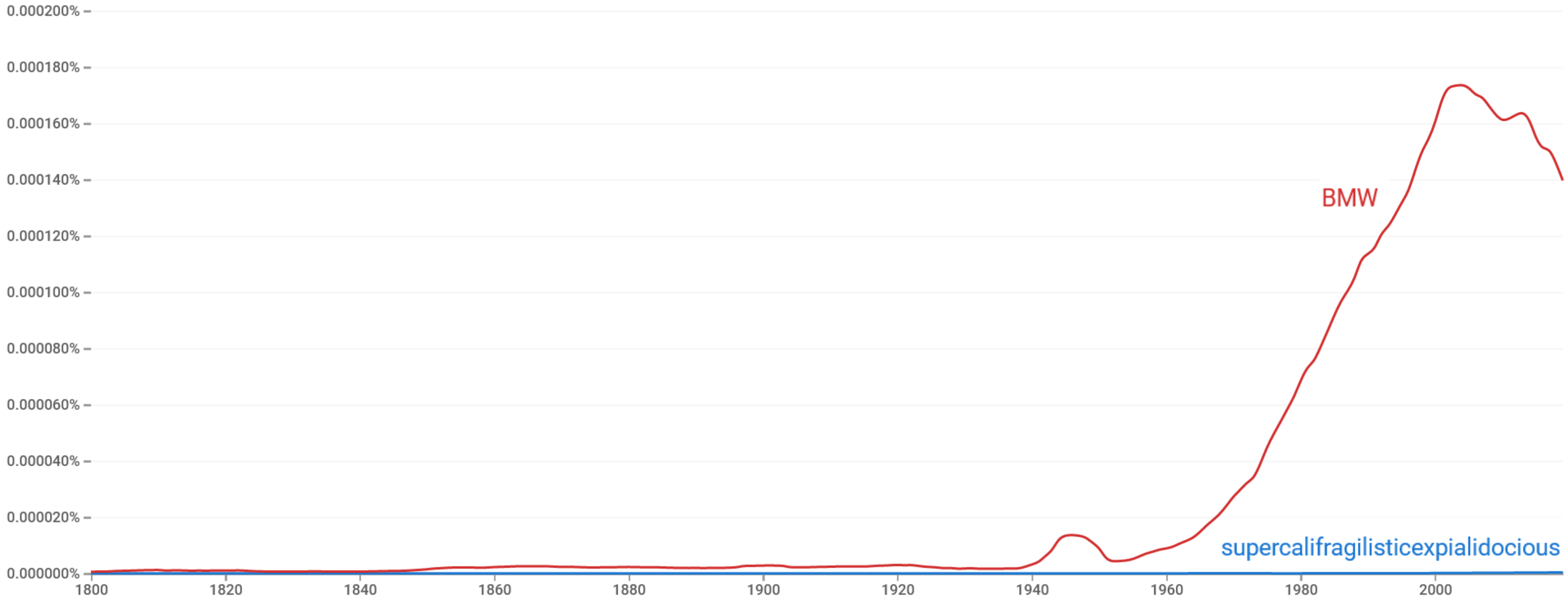


Zipf‘s Law



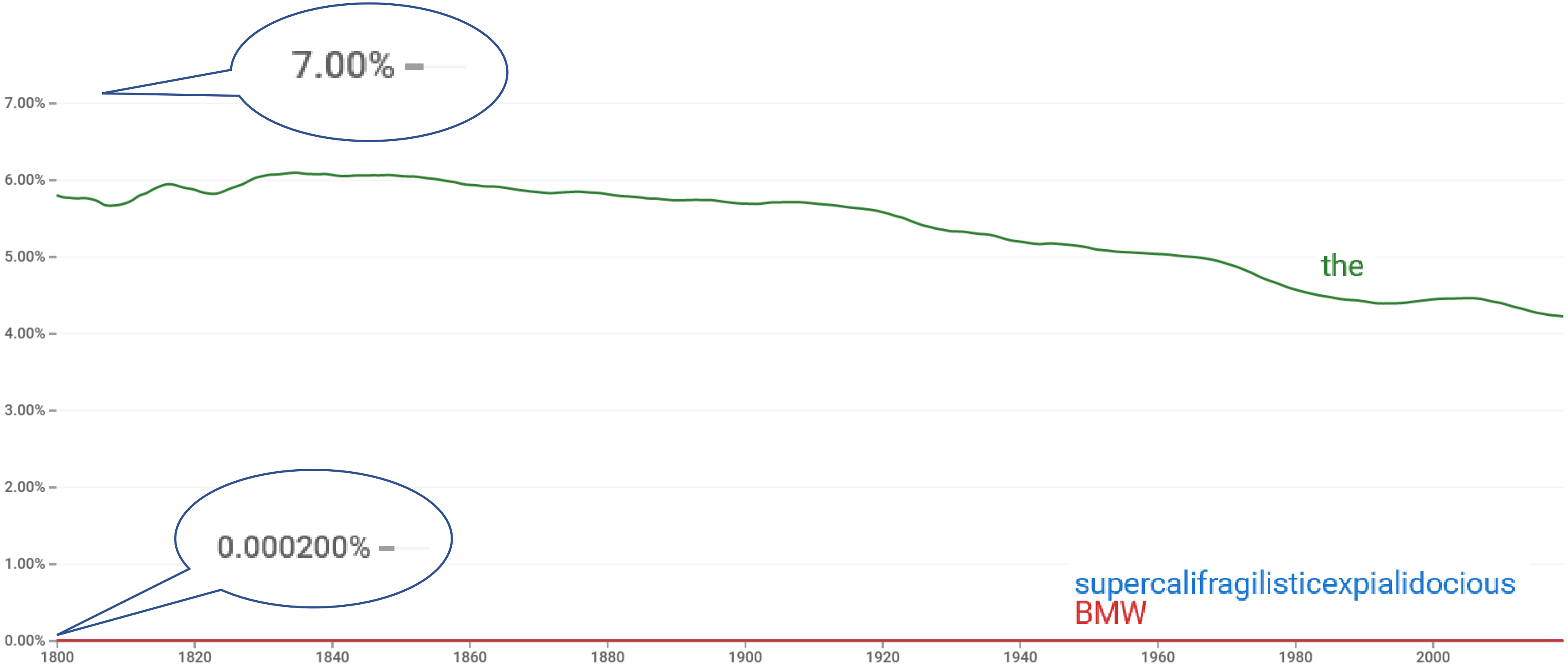
http://wugology.com/zipfs-law/

What does it mean?



https://books.google.com/ngrams

What does it mean?



https://books.google.com/ngrams

If data sparsity isn’t a problem for you, your model is too simple!

Chen & Goodman (1998)

If data sparsity isn't a problem for you, your model is too simple!

"Whenever data sparsity is an issue, smoothing can help performance, and data sparsity is almost always an issue in statistical modeling. In the extreme case where there is so much training data that all parameters can be accurately trained without smoothing, one can almost always expand the model, such as by moving to a higher n-gram model, to achieve improved performance. With more parameters data sparsity becomes an issue again, but with proper smoothing the models are usually more accurate than the original models. Thus, no matter how much data one has, smoothing can almost always help performance, and for a relatively small effort."

Chen & Goodman (1998)

Repeat

•JOHN READ MOBY DICK•
•MARY READ A DIFFERENT BOOK•
•SHE READ A BOOK BY CHER•

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_{w_i} c(w_{i-1}w_i)}$$

$$p(s) = \prod_{i=1}^{l+1} p(w_i|w_{i-1})$$

Example

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$
\begin{aligned}
& p(\bullet\,\text{JOHN READ A BOOK}\,\bullet) \\
= \quad & p(\text{JOHN} \mid \bullet) \\
= \quad & \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \\
= \quad & ?
\end{aligned}
$$

Example

$\bullet$ JOHN READ MOBY DICK $\bullet$
$\bullet$ MARY READ A DIFFERENT BOOK $\bullet$
$\bullet$ SHE READ A BOOK BY CHER $\bullet$

$$p(\bullet\ \text{JOHN READ A BOOK}\ \bullet)$$

$$= \quad p(\text{JOHN} \mid \bullet)$$

$$= \quad \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)}$$

$$= \quad \frac{1}{3}$$

Technische Hochschule Ingolstadt

Example

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$p(\bullet\text{JOHN READ A BOOK}\bullet)$$

$$= p(\text{JOHN}|\bullet) \quad p(\text{READ}|\text{JOHN})$$

$$= \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN}\ w)}$$

$$= \frac{1}{3} \quad ?$$

Example

$\bullet$ JOHN READ MOBY DICK $\bullet$
$\bullet$ MARY READ A DIFFERENT BOOK $\bullet$
$\bullet$ SHE READ A BOOK BY CHER $\bullet$

$$
\begin{aligned}
& p(\bullet \text{JOHN READ A BOOK} \bullet) \\
= \quad & p(\text{JOHN} \mid \bullet) \quad p(\text{READ} \mid \text{JOHN}) \\
= \quad & \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN}\ w)} \\
= \quad & \frac{1}{3} \qquad\qquad \frac{1}{1}
\end{aligned}
$$

Example

$$\bullet \text{ JOHN READ MOBY DICK } \bullet$$
$$\bullet \text{ MARY READ A DIFFERENT BOOK } \bullet$$
$$\bullet \text{ SHE READ A BOOK BY CHER } \bullet$$

$$
\begin{aligned}
& p(\bullet \text{JOHN READ A BOOK} \bullet) \\
= \;& p(\text{JOHN}|\bullet) \quad p(\text{READ}|\text{JOHN}) \quad p(\text{A}|\text{READ}) \\
= \;& \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN}\ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)} \\
= \;& \frac{1}{3} \quad \frac{1}{1} \quad ?
\end{aligned}
$$

Example

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$p(\bullet \text{JOHN READ A BOOK} \bullet)$$

$$= \; p(\text{JOHN}|\bullet) \quad p(\text{READ}|\text{JOHN}) \quad p(\text{A}|\text{READ})$$

$$= \; \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN}\ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)}$$

$$= \; \frac{1}{3} \quad \frac{1}{1} \quad \frac{2}{3}$$

Example

$$\bullet \text{JOHN READ MOBY DICK} \bullet$$
$$\bullet \text{MARY READ A DIFFERENT BOOK} \bullet$$
$$\bullet \text{SHE READ A BOOK BY CHER} \bullet$$

$$p(\bullet \text{JOHN READ A BOOK} \bullet)$$

$$= \quad p(\text{JOHN}|\bullet) \quad p(\text{READ}|\text{JOHN}) \quad p(\text{A}|\text{READ}) \quad p(\text{BOOK}|\text{A})$$

$$= \quad \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN}\ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)} \quad \frac{c(\text{A BOOK})}{\sum_w c(\text{A}\ w)}$$

$$= \quad \frac{1}{3} \quad \frac{1}{1} \quad \frac{2}{3} \quad ?$$

Example

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$
\begin{aligned}
& p(\bullet\,\text{JOHN READ A BOOK}\,\bullet) \\
= \;& p(\text{JOHN}|\bullet) \quad p(\text{READ}|\text{JOHN}) \quad p(\text{A}|\text{READ}) \quad p(\text{BOOK}|\text{A}) \\
= \;& \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN}\ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)} \quad \frac{c(\text{A BOOK})}{\sum_w c(\text{A}\ w)} \\
= \;& \frac{1}{3} \quad \frac{1}{1} \quad \frac{2}{3} \quad \frac{1}{2}
\end{aligned}
$$

Technische Hochschule Ingolstadt

Example

$$\bullet \text{JOHN READ MOBY DICK} \bullet$$
$$\bullet \text{MARY READ A DIFFERENT BOOK} \bullet$$
$$\bullet \text{SHE READ A BOOK BY CHER} \bullet$$

$$p(\bullet \text{JOHN READ A BOOK} \bullet)$$

$$= p(\text{JOHN}|\bullet) \; p(\text{READ}|\text{JOHN}) \; p(\text{A}|\text{READ}) \; p(\text{BOOK}|\text{A}) \; p(\bullet|\text{BOOK})$$

$$= \frac{c(\bullet\ \text{JOHN})}{\sum_w c(\bullet\ w)} \; \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN}\ w)} \; \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)} \; \frac{c(\text{A BOOK})}{\sum_w c(\text{A}\ w)} \; \frac{c(\text{BOOK}\ \bullet)}{\sum_w c(\text{BOOK}\ w)}$$

$$= \frac{1}{3} \; \frac{1}{1} \; \frac{2}{3} \; \frac{1}{2} \; ?$$

Example

$$\bullet \text{ JOHN READ MOBY DICK } \bullet$$
$$\bullet \text{ MARY READ A DIFFERENT BOOK } \bullet$$
$$\bullet \text{ SHE READ A BOOK BY CHER } \bullet$$

$$p(\bullet \text{JOHN READ A BOOK} \bullet)$$

$$= \; p(\text{JOHN}|\bullet) \; p(\text{READ}|\text{JOHN}) \; p(\text{A}|\text{READ}) \; p(\text{BOOK}|\text{A}) \; p(\bullet|\text{BOOK})$$

$$= \; \frac{c(\bullet \text{ JOHN})}{\sum_w c(\bullet \; w)} \; \frac{c(\text{JOHN READ})}{\sum_w c(\text{JOHN } w)} \; \frac{c(\text{READ A})}{\sum_w c(\text{READ } w)} \; \frac{c(\text{A BOOK})}{\sum_w c(\text{A } w)} \; \frac{c(\text{BOOK } \bullet)}{\sum_w c(\text{BOOK } w)}$$

$$= \; \frac{1}{3} \; \frac{1}{1} \; \frac{2}{3} \; \frac{1}{2} \; \frac{1}{2}$$

$$\approx \; 0.06$$

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

p(• JOHN READ A BOOK •)

p(• CHER READ A BOOK •)

- Data for training remains the same.
- Data in the test changes.

Example: How well does it generalize?

- $\bullet$ JOHN READ MOBY DICK $\bullet$
- $\bullet$ MARY READ A DIFFERENT BOOK $\bullet$
- $\bullet$ SHE READ A BOOK BY CHER $\bullet$

$p(\bullet\ \text{CHER READ A BOOK}\ \bullet)$

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$p(\bullet\ \text{CHER READ A BOOK}\ \bullet)$$

$$= \quad p(\text{CHER} \mid \bullet)$$

$$= \quad \frac{c(\bullet\ \text{CHER})}{\sum_w c(\bullet\ w)}$$

$$= \quad ?$$

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$
\begin{aligned}
& p(\bullet\,\text{CHER READ A BOOK}\,\bullet) \\
= \;& p(\text{CHER} \mid \bullet) \\
= \;& \frac{c(\bullet\ \text{CHER})}{\sum_w c(\bullet\ w)} \\
= \;& \frac{0}{3}
\end{aligned}
$$

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$p(\bullet \text{ CHER READ A BOOK } \bullet)$$

$$= p(\text{CHER}|\bullet) \quad p(\text{READ}|\text{CHER})$$

$$= \frac{c(\bullet \text{ CHER})}{\sum_w c(\bullet \; w)} \quad \frac{c(\text{CHER READ})}{\sum_w c(\text{CHER } w)}$$

$$= \frac{0}{3} \qquad ?$$

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$p(\bullet \text{CHER READ A BOOK} \bullet)$$

$$= \; p(\text{CHER}|\bullet) \quad p(\text{READ}|\text{CHER})$$

$$= \; \frac{c(\bullet\ \text{CHER})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{CHER READ})}{\sum_w c(\text{CHER}\ w)}$$

$$= \; \frac{0}{3} \quad \frac{0}{1}$$

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$p(\bullet \text{CHER READ A BOOK} \bullet)$$

$$= p(\text{CHER}|\bullet) \quad p(\text{READ}|\text{CHER}) \quad p(\text{A}|\text{READ})$$

$$= \frac{c(\bullet \ \text{CHER})}{\sum_w c(\bullet \ w)} \quad \frac{c(\text{CHER READ})}{\sum_w c(\text{CHER} \ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ} \ w)}$$

$$= \frac{0}{3} \quad \frac{0}{1} \quad ?$$

Technische Hochschule Ingolstadt

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

$$p(\bullet\,\text{CHER READ A BOOK}\,\bullet)$$

$$= p(\text{CHER}|\bullet) \quad p(\text{READ}|\text{CHER}) \quad p(\text{A}|\text{READ})$$

$$= \frac{c(\bullet\ \text{CHER})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{CHER READ})}{\sum_w c(\text{CHER}\ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)}$$

$$= \frac{0}{3} \quad \frac{0}{1} \quad \frac{2}{3}$$

Example: How well does it generalize?

$\bullet$ JOHN READ MOBY DICK $\bullet$
$\bullet$ MARY READ A DIFFERENT BOOK $\bullet$
$\bullet$ SHE READ A BOOK BY CHER $\bullet$

$$
\begin{aligned}
& p(\bullet \text{CHER READ A BOOK} \bullet) \\
= \; & p(\text{CHER}|\bullet) \quad p(\text{READ}|\text{CHER}) \quad p(\text{A}|\text{READ}) \quad p(\text{BOOK}|\text{A}) \quad p(\bullet|\text{BOOK}) \\
= \; & \frac{c(\bullet\ \text{CHER})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{CHER READ})}{\sum_w c(\text{CHER}\ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)} \quad \frac{c(\text{A BOOK})}{\sum_w c(\text{A}\ w)} \quad \frac{c(\text{BOOK}\ \bullet)}{\sum_w c(\text{BOOK}\ w)} \\
= \; & \frac{0}{3} \quad \frac{0}{1} \quad \frac{2}{3} \quad \frac{1}{2} \quad \frac{1}{2}
\end{aligned}
$$

Technische Hochschule Ingolstadt

Example: How well does it generalize?

• JOHN READ MOBY DICK •
• MARY READ A DIFFERENT BOOK •
• SHE READ A BOOK BY CHER •

Does not occur in training!

$$p(\bullet\,\text{CHER READ A BOOK}\,\bullet)$$

$$= p(\text{CHER}|\bullet) \quad p(\text{READ}|\text{CHER}) \quad p(\text{A}|\text{READ}) \quad p(\text{BOOK}|\text{A}) \quad p(\bullet|\text{BOOK})$$

$$= \frac{c(\bullet\ \text{CHER})}{\sum_w c(\bullet\ w)} \quad \frac{c(\text{CHER READ})}{\sum_w c(\text{CHER}\ w)} \quad \frac{c(\text{READ A})}{\sum_w c(\text{READ}\ w)} \quad \frac{c(\text{A BOOK})}{\sum_w c(\text{A}\ w)} \quad \frac{c(\text{BOOK}\ \bullet)}{\sum_w c(\text{BOOK}\ w)}$$

$$= \frac{0}{3} \quad \frac{0}{1} \quad \frac{2}{3} \quad \frac{1}{2} \quad \frac{1}{2}$$

$$= 0$$

How to solve the problem?

?

How to solve the problem?

- Augment
  - with real data from other sources
  - with simulated data from expert knowledge

Flight Air Control Example
„Manching Tower: CityAirbus 007, Wind 230 degrees, 5 knots, cleared for take-off"

=> (Let's discuss soon.)

- What other options are there?

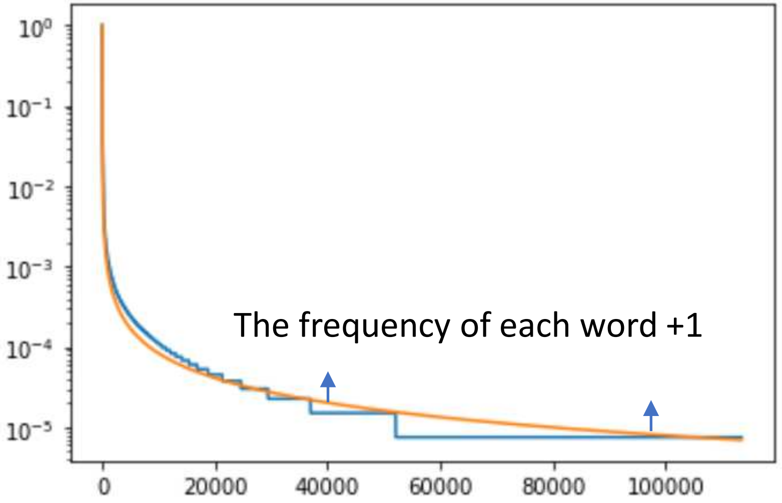
How to solve the problem?

- Augment
  - with real data from other sources
  - with simulated data from expert knowledge

Flight Air Control Example
„Manching Tower: CityAirbus 007, Wind 230 degrees, 5 knots, cleared for take-off"

- Smooth
- Interpolate
- Smoothing & Interpolating

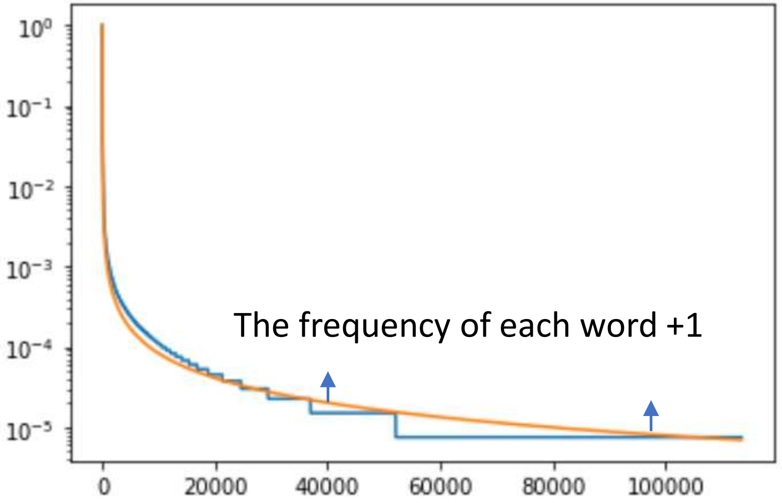
Laplace (add-one) Smoothing



We remember:

$$p(w_i|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_{w_i} c(w_{i-1}w_i)}$$

Laplace (add-one) Smoothing



$$p(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

with

$$V = \{w : c(w) > 0\} \cup \{\text{UNK}\}$$

Technische Hochschule Ingolstadt

Laplace (add-one) Smoothing

JOHN READ MOBY DICK
MARY READ A DIFFERENT BOOK
SHE READ A BOOK BY CHER

$p(\text{JOHN READ A BOOK})$

?

$p(\text{CHER READ A BOOK})$

?

$$p(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

with

$$V = \{w : c(w) > 0\} \cup \{\text{UNK}\}$$

Laplace (add-one) Smoothing

JOHN READ MOBY DICK
MARY READ A DIFFERENT BOOK
SHE READ A BOOK BY CHER

$p(\text{JOHN READ A BOOK})$

$= \frac{1+1}{11+3} \; \frac{1+1}{11+1} \; \frac{1+2}{11+3} \; \frac{1+1}{11+2} \; \frac{1+1}{11+2}$

$\approx \; 0.0001$

$p(\text{CHER READ A BOOK})$

?

$$p(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

with

$$V = \{w : c(w) > 0\} \cup \{\text{UNK}\}$$

Technische Hochschule Ingolstadt

Laplace (add-one) Smoothing

JOHN READ MOBY DICK
MARY READ A DIFFERENT BOOK
SHE READ A BOOK BY CHER

$$p(\text{JOHN READ A BOOK})$$

$$= \frac{1+1}{11+3} \; \frac{1+1}{11+1} \; \frac{1+2}{11+3} \; \frac{1+1}{11+2} \; \frac{1+1}{11+2}$$

$$\approx \; 0.0001$$

$$p(\text{CHER READ A BOOK})$$

$$= \frac{1+0}{11+3} \; \frac{1+0}{11+1} \; \frac{1+2}{11+3} \; \frac{1+1}{11+2} \; \frac{1+1}{11+2}$$

$$\approx \; 0.00003$$

$$p(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

with

$$V = \{w : c(w) > 0\} \cup \{\text{UNK}\}$$

Works better already.

But that can be done even better!

Technische Hochschule Ingolstadt

Additive Smoothing

- Add a constant to each n-gram frequency
- Lidstone & Jeffreys: $\delta = 1$
  „add-one“ Smoothing

$$p(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

with

$$V = \{w : c(w) > 0\} \cup \{\text{UNK}\}$$

Generalization

$$p_{add}(w_i|w_{i-n+1}^{i-1}) = \frac{\delta + c(w_{i-n+1}^{i})}{\delta|V| + \sum_{w_i} c(w_{i-n+1}^{i})}$$

Additive Smoothing

- Add a constant to each n-gram frequency
- Laplace/„add-one“: $\delta = 1$
- Lidstone & Jeffreys:

  “A large dictionary makes novel events too probable”
- Common $0 < \delta \leq 1$
- How to determine? δ

$$p(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

with

$$V = \{w : c(w) > 0\} \cup \{\text{UNK}\}$$

Generalization

$$p_{add}(w_i|w_{i-n+1}^{i-1}) = \frac{\delta + c(w_{i-n+1}^{i})}{\delta|V| + \sum_{w_i} c(w_{i-n+1}^{i})}$$

Additive Smoothing

- Add a constant to each n-gram frequency
- Laplace/„add-one“: $\delta = 1$
- Lidstone & Jeffreys:

  “A large dictionary makes novel events too probable”
- Common $0 < \delta \leq 1$
- How to determine δ ?

⇩

Test, Test, Test, ….

$$p(w_i|w_{i-1}) = \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i}[1 + c(w_{i-1}w_i)]}$$

$$= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)}$$

with

$$V = \{w : c(w) > 0\} \cup \{\mathsf{UNK}\}$$

Generalization

$$p_{add}(w_i|w_{i-n+1}^{i-1}) = \frac{\delta + c(w_{i-n+1}^{i})}{\delta|V| + \sum_{w_i} c(w_{i-n+1}^{i})}$$

Technische Hochschule Ingolstadt

Feynman’s Advice:

„The first principle is that

you must not fool yourself,

and

you are the easiest person to fool.“

Always test thoroughly when setting hyperparameters.Always test on at least two data sets
(Better: train-, held out-, dev-, test-, validation-data).

https://de.wikipedia.org/wiki/Richard_Feynman