自然言語処理シリーズ I

言語モデルのための
情報理論入門

岡　照晃=監修

岡　照晃=著

Natural Language Processing Series

参考文献


情報理論
甘利俊一
あくまで学芸文庫


自然言語処理シリーズ 1
言語処理のための
機械学習入門
奥村 学=監修
Natural Language Processing Series
コロナ社


辻井潤一=編
言語と計算 4
確率的言語モデル
北 研二
東京大学出版会

エントロピー編

配点： それがなんだか一言で言えますか？

エントロピー

- **状況の不確定度を表す量**
  - 不確定度が高いと値が大きくなり，不確定度が小さいと値が小さくなる.

$$H(X) = -\sum_{x \in X} p(x)\, log_2\, p(x) \quad \geqq 0 \quad \text{[bit]}$$

X: 標本空間（全事象）

例えば，普通のサイコロ

- **各目の出る確率は同じ（一様分布）．**
  - X={1, 2, 3, 4, 5, 6}
  - p(x)=1/6（xに依らない）

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

$$= -6\left(\frac{1}{6} log_2 \frac{1}{6}\right) \fallingdotseq 2.58$$

対して，必ず１の目が出るとわかっているイカサマサイコロ

- $p(1) = 1$
- $p(x) = 0 \quad (x \neq 1)$

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

$$= -(1\, log_2\, 1) - 5(0\, log_2\, 0) = 0$$

（但し，$0 \log_2 0 = 0$ ）

- **不確定度が低いと，エントロピーも低い！**

ちなみに，エントロピーが一番大きくなるのは一様分布

- **証明：**

$$H = -\sum p_i \log p_i$$

を最大にする分布$p_1, p_2, \ldots, p_i, \ldots$を求める.

ただし，　$\sum p_i = 1$

次ページへ続く☞

みんな大好きラグランジュの未定乗数法

ここではlogの底はeとする

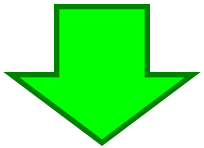
$H + \lambda\left(1 - \sum p_i\right)$ を各p_iで微分して0とおく.

$$\frac{\partial}{\partial p_i}\left\{\left(-\sum p_i \log p_i\right) + \lambda - \lambda \sum p_i\right\} = -\log p_i - 1 - \lambda = 0$$

$$\log p_i = -1 - \lambda$$

$$p_i = e^{-1-\lambda}$$ ← i に依らない定数 = **一様分布**

- ここまでのお話は，事象がただ一回だけ独立に起こって，それでおしまいとなる場合.
- しかし，実際には，これから起こる事象というのは過去に起こった事象に依存することが多い.

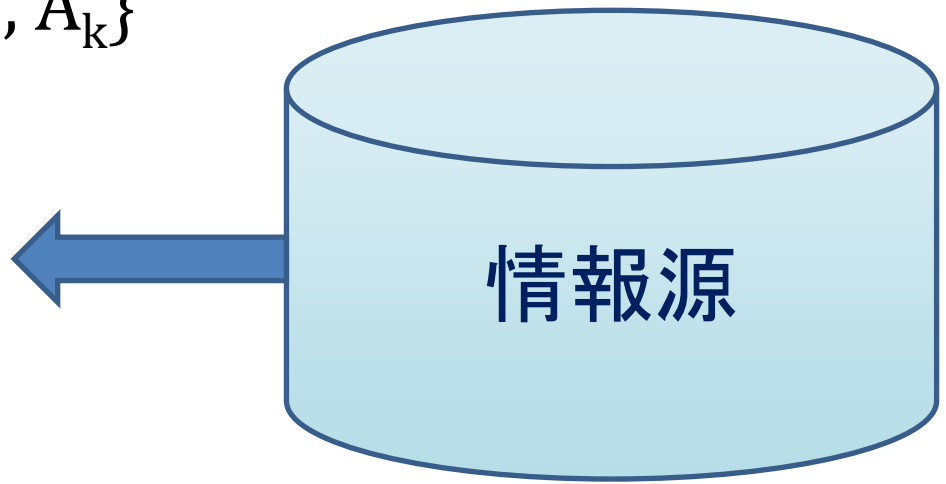


情報源という考え方を導入

情報源

全事象：$X=\{A_1, A_2, \ldots, A_k\}$

……$A_3A_1A_2A_5A_6$……

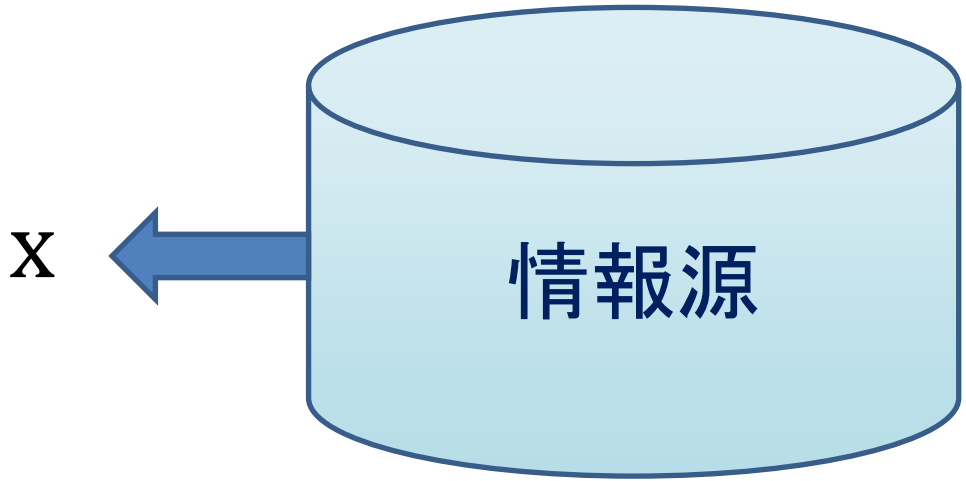


情報源：

- 次から次へと無限に有限種類の事象を生成する.
- これから出てくる事象は過去の事象に依存し，確率的に定まる.
  （現在の事象の確率分布は過去に発生した事象がなんであったかに影響される）

情報源のエントロピー

- 情報源から事象が１個現れるときのエントロピー.



そう言われると，これの気がするが…

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

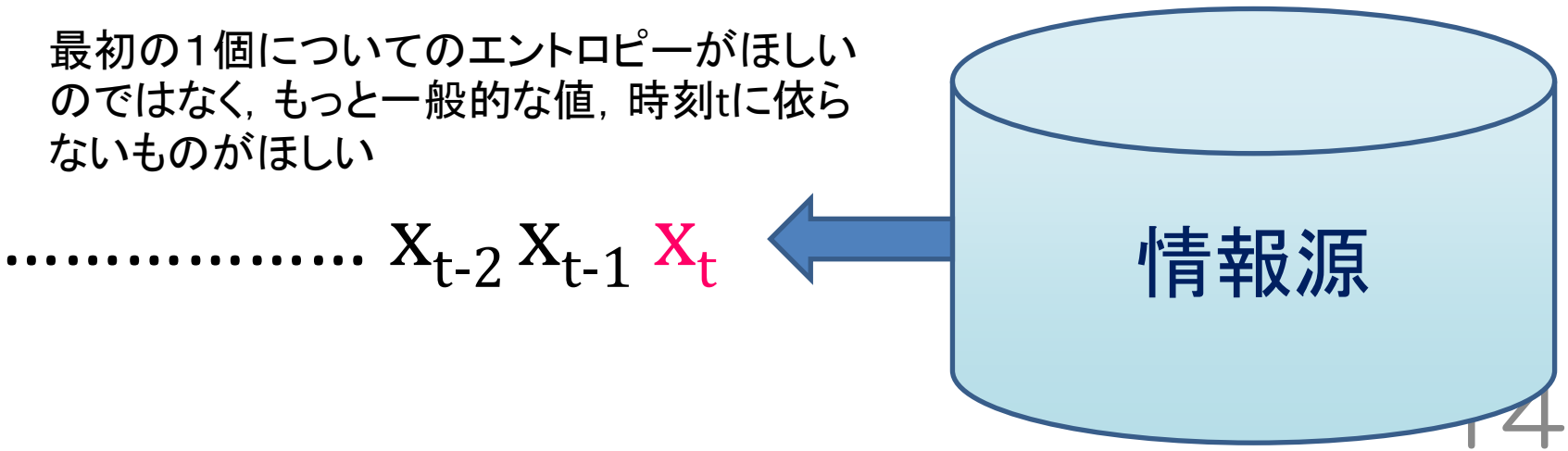
そう言われると，これの気がするが…

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

情報源のエントロピー

- 情報源は事象を生み出し続けているので，
  x_t が現れるときは，x_{t-1} や x_{t-2} を含め，
  それらよりずっと前の事象にも影響を受け得る．

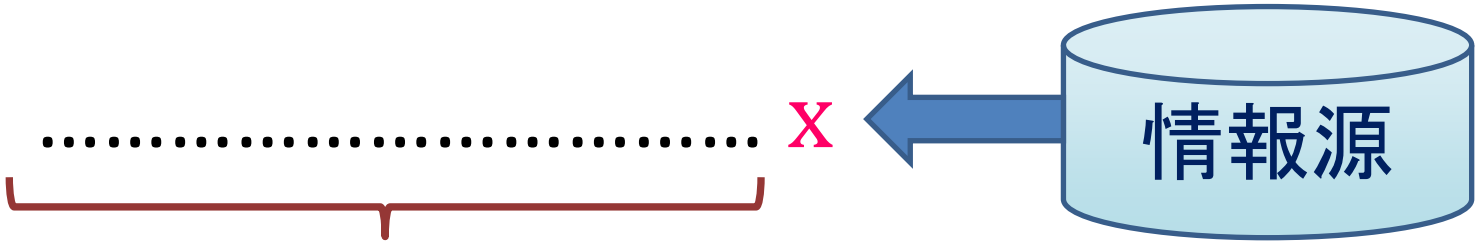
最初の１個についてのエントロピーがほしいのではなく，もっと一般的な値，時刻tに依らないものがほしい



だからこう

$$H(X|X^{\infty}) = \lim_{n\to\infty} H(X|X^{n})$$

依存関係を考慮するのに十分に長い（無限長）の事象の列を確認した上で，次に出て来る事象の不確定度.



ここで条件付け（期待値をとるから具体的でなくてもいい）

条件付エントロピー

$$H(X|Y) = -\sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) \log p(x|y)$$

y固定

yについての期待値をとる

但し，各事象が互いに独立であれば
こちらでもいい

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

条件付エントロピーは計算が面倒なので

- **エントロピーの展開:**

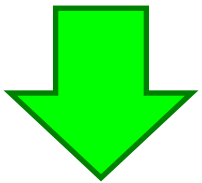
$$H(X^n) = H(X) + H(X|X) + H(X|X^2) + \cdots + H(X|X^{n-1})$$

$$H(X|X^\infty) = \lim_{n\to\infty} \frac{H(X^n)}{n} = H$$

条件付きエントロピーを複合事象のエントロピーで表す

$$H(X^n) = -\sum_{x_1{}^n} p(x_1{}^n) \log p(x_1{}^n)$$ より，

$$H = -\lim_{n\to\infty} \frac{1}{n} \sum_{x_1{}^n} p(x_1{}^n) \log p(x_1{}^n)$$



定常性とエルゴート性を仮定すれば

Shannon-McMillan-Breiman theorem により，

（証明は，それだけで論文1本書ける内容なので割愛）

[Algoet and Cover, 1998]

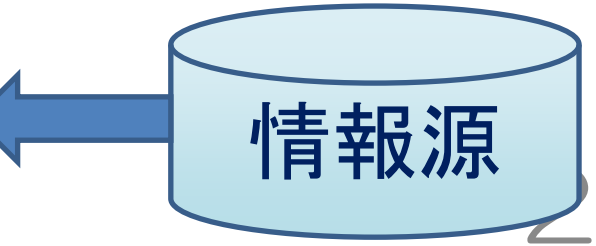
$$H = -\lim_{n\to\infty} \frac{1}{n} \log p(x_1{}^n)$$

まとめると，

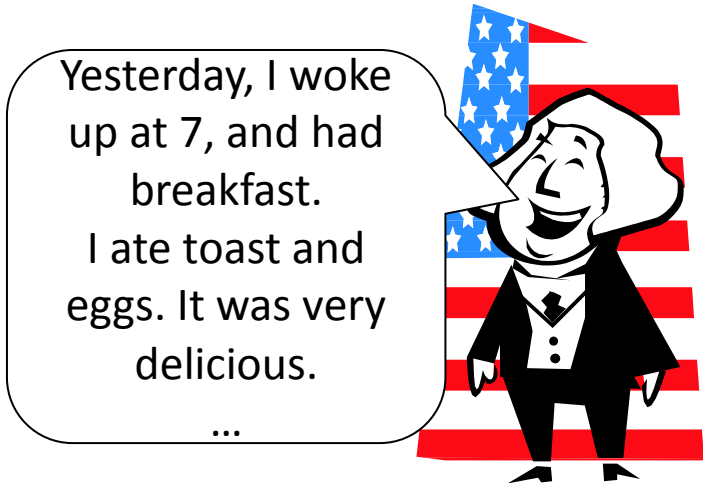
- **情報源のエントロピー H：**
  - 情報源から事象が１個現れるときのエントロピー．
    - 依存関係を考慮するのに十分に長い（無限長）の事象の列を確認した上で，次に出て来る事象の不確定度．

$$H = -\lim_{n\to\infty} \frac{1}{n} \log p(x_1{}^n)$$

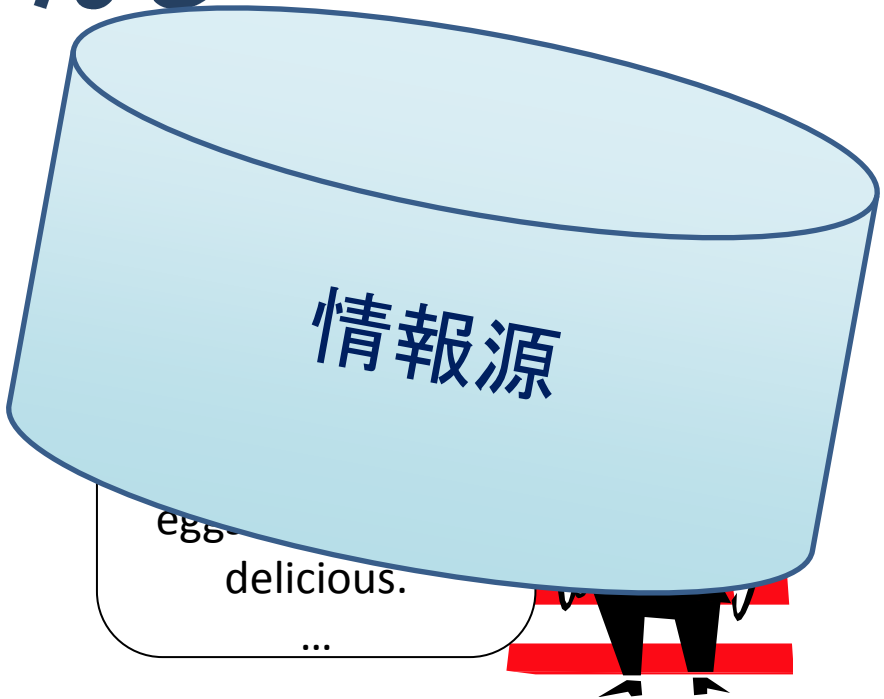
…… x



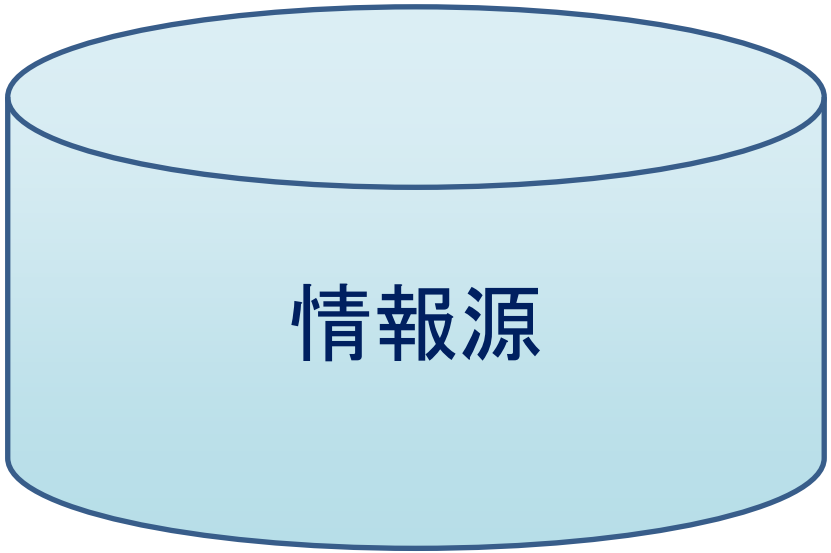
情報源を言語モデルの文脈でいうと，こんなかんじ



情報源を言語モデルの文脈でいうと，こんなかんじ

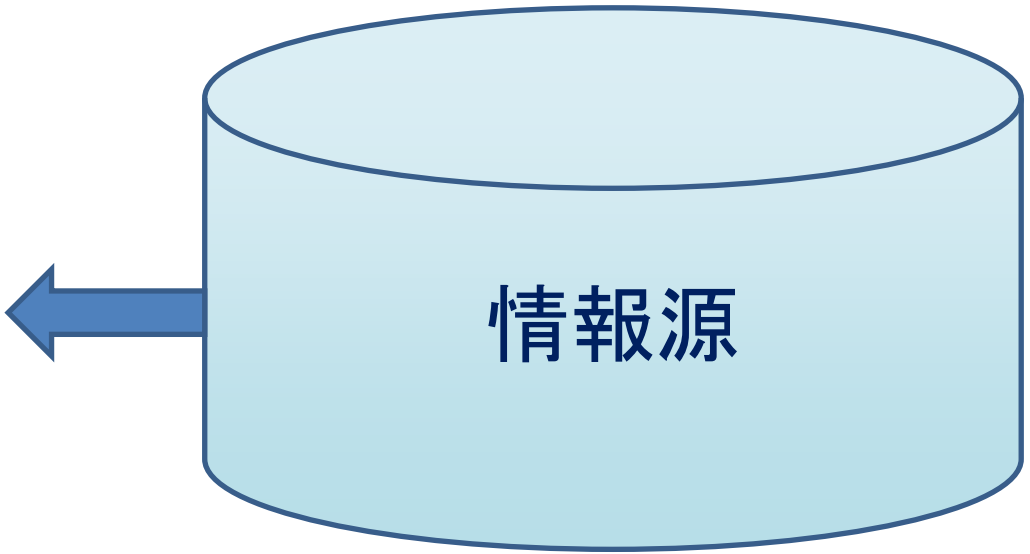


情報源を言語モデルの文脈でいうと，こんなかんじ

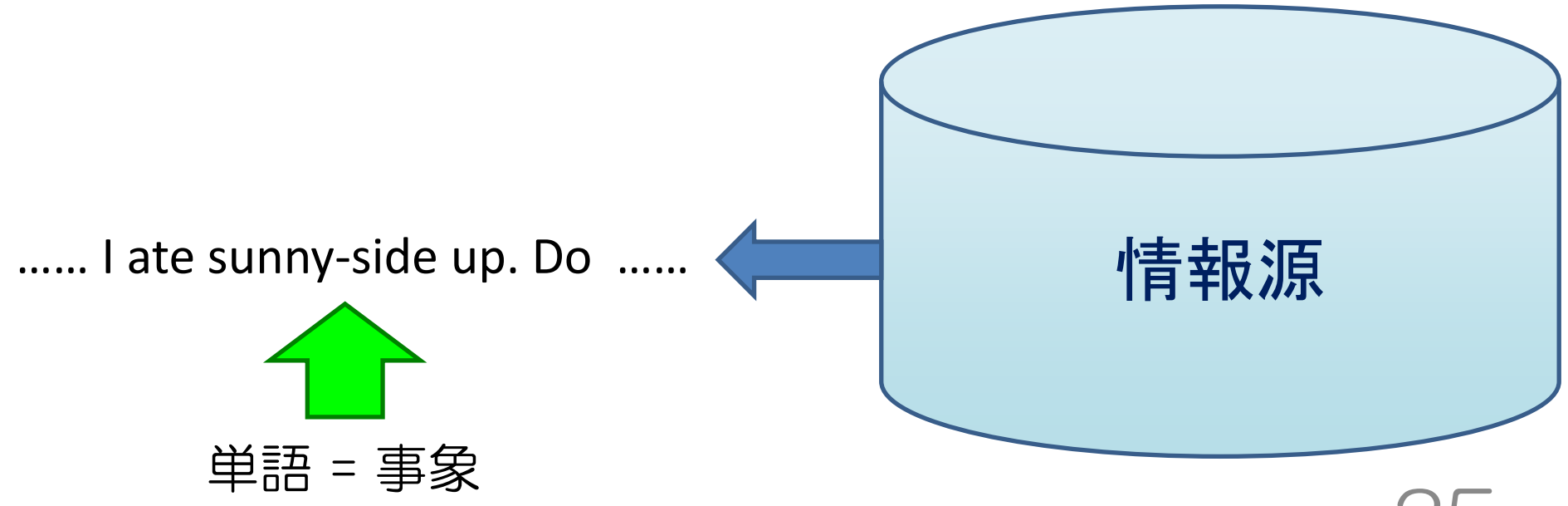


情報源を言語モデルの文脈でいうと，こんなかんじ

…… I ate sunny-side up. Do ……

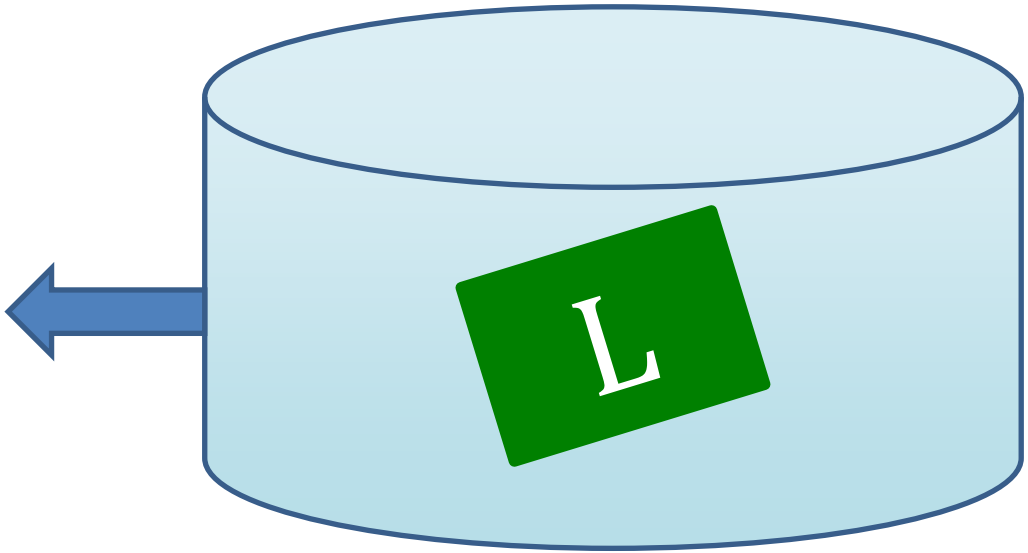


情報源を言語モデルの文脈でいうと，こんなかんじ



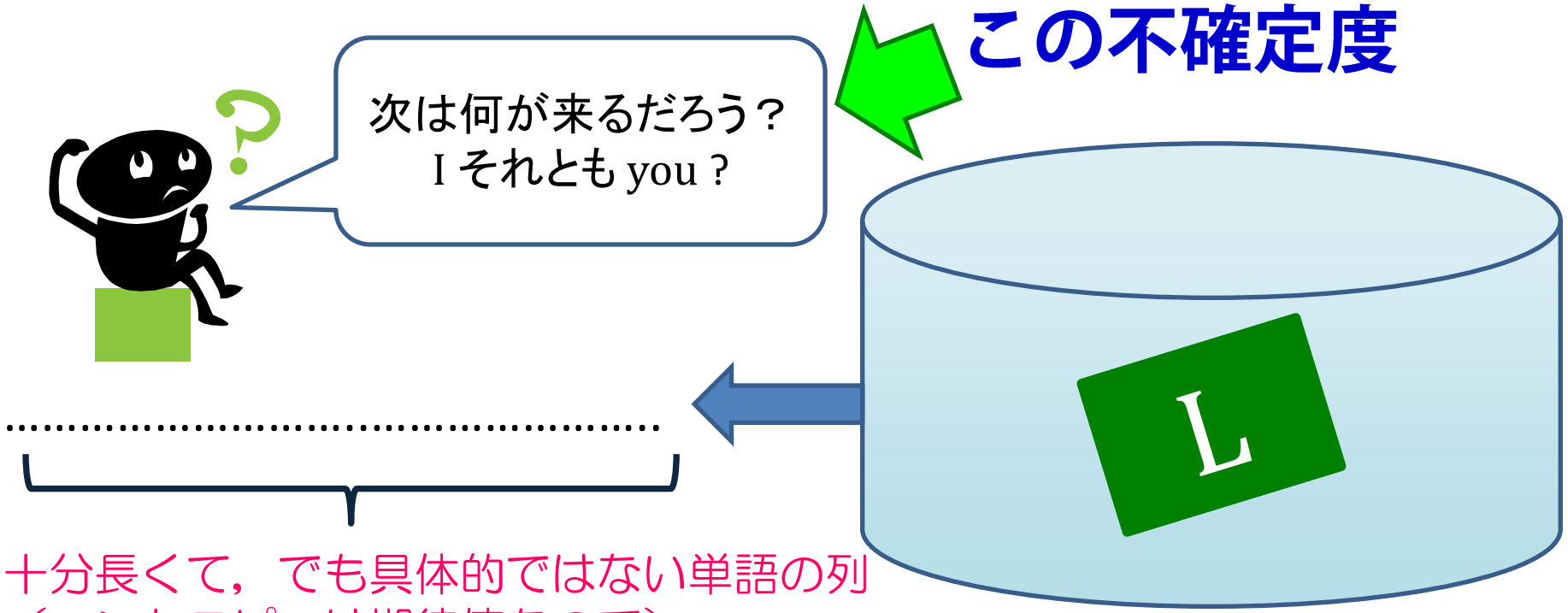
情報源を言語モデルの文脈でいうと，こんなかんじ

…… I ate sunny-side up. Do ……



単語 = 事象

言語Lのエントロピー



十分長くて，でも具体的ではない単語の列
（エントロピーは期待値なので）

式で書くと

$$H(L) = \lim_{n\to\infty} -\frac{1}{n}\log P_L(w_1 w_2 \ldots w_n)$$

ポイント：　ここまでの話の中に，「１つ前を見るだけでいい」とか「マルコフ性」とか「Ngram」とかは一切出てきていない．

今考えている$P_L(w_1 w_2 \ldots w_n)$は，真の言語モデルであり，N-gramを使って計算できるのはその近似でしかない．

ここ，混同しないように注意

コラム：定常性とエルゴート性

- **定常性：**
  - 情報源の性質が時間がたっても変わらないということ.
  - 日本語だけを出していた情報源があるときを境に英語しか出さなくなるとかそういうことがないこと.
    - 例えば，時刻tにおいて，$P_L(\text{study}) = 0.1$だったのが，時刻t+rで0.5になったりはしないということ.

情報源から出て来る１個目の単語の確率

- **エルゴート性：**
  - １本の十分に長い事象列を観測すれば，その中に他の事象列が全部含まれているというカンジの性質.
  - 集団平均＝時間平均が成り立つ.

パープレキシティー編

配点： 真の言語モデルをどれだけ近似できてるだろう？

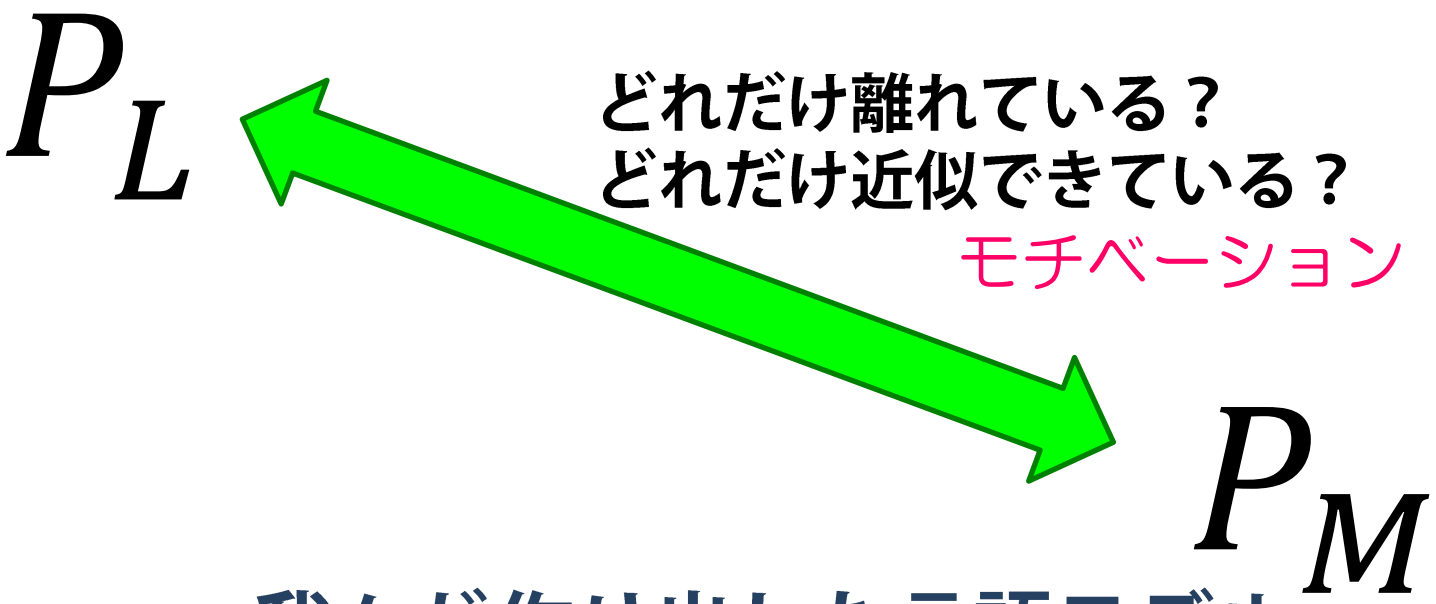
KLダイバージェンス

- ２つの確率分布間の距離（異なり度）．

$$D_{KL}(P||Q) = \sum_{x} P(X = x) \log \frac{P(X = x)}{Q(X = x)}$$

$D_{KL} \geq 0$ （ D_{KL} =0になるのはP=Qの時に限る）

我々がたどり着くべき真の言語モデル
(N-gramとかで近似とかしない本物)



我々が作り出した言語モデル
(N-gramモデルとか)

KLダイバージェンスで測ってみよう

$$D_{KL}(P_L||P_M) = \lim_{n\to\infty} \sum_{{w_1}^n} P_L({w_1}^n) \log \frac{P_L({w_1}^n)}{P_M\ ({w_1}^n)}$$

$$= \left[-\lim_{n\to\infty} \sum_{{w_1}^n} P_L({w_1}^n) \log P_M\ ({w_1}^n) \right] - \left[-\lim_{n\to\infty} \sum_{{w_1}^n} P_L({w_1}^n) \log P_L\ ({w_1}^n) \right] \geq 0$$

n>0で割ると，

$$\left[-\lim_{n\to\infty}\frac{1}{n}\sum_{x_1{}^n} P_L(w_1{}^n)\log P_M\ (w_1{}^n)\right]-\left[-\lim_{n\to\infty}\frac{1}{n}\sum_{x_1{}^n} P_L(w_1{}^n)\log P_L\ (w_1{}^n)\right]\geq 0$$

新登場

クロスエントロピー

$\mathrm{H}(\mathrm{P_L}, \mathrm{P_M})$

言語Lのエントロピー

$\mathrm{H}(\mathrm{L})$

エントロピーは0よりも大きいから，クロスエントロピーは言語Lのエントロピーのアッパーバウンドになっている．

$\mathrm{H}(\mathrm{P_L}, \mathrm{P_M}) \geqq \mathrm{H}(\mathrm{L})$

エントロピーは0よりも大きいから，クロスエントロピーは言語Lのエントロピーのアッパーバウンドになっている.

$H(P_L, P_M) \geqq H(L)$

$H(L)$は不変だから，クロスエントロピーをより小さくできれば，より真の言語モデルに近い（KLダイバージェンスが低い）モデルが作れたことになる.

意味合い的には，クロスエントロピー≒真の言語モデルまでの距離

必殺，シャノン・マクミラン・ブレイマン！

$$H(P_L, P_M) = -\lim_{n\to\infty} \frac{1}{n} \sum_{w_1{}^n} P_L(w_1{}^n) \log P_M\ (w_1{}^n)$$

$$H(P_L, P_M) = -\lim_{n\to\infty} \frac{1}{n} \log P_M\ (w_1{}^n)$$

n→∞とか無理だし，
適当に大きな値で妥協

$$H(P_L, P_M) = -\lim_{n\to\infty} \frac{1}{n} \log P_M\left(w_1{}^n\right)$$

$$H(\text{評価用コーパス}) = -\frac{1}{n} \log P_M\left(w_1{}^n\right)$$

$w_1{}^n$： 評価用コーパスから作った１本の単語列.
（長ければ長いほど本来の値（n=∞）に近づく）

パープレキシティ

$$H(\text{評価用コーパス}) = -\frac{1}{n}\log P_M\,(w_1{}^n)$$

N-gram言語モデル

$$Perplexty(\text{評価用コーパス}) = 2^{H(\text{評価用コーパス})}$$

$$= P_M(w_1 w_2 \ldots w_n)^{-\frac{1}{n}}$$

つまり，

パープレキシティ

≒ クロスエントロピー

≒ 真の言語モデルのまでの距離

（KLダイバージェンス）

別の見方をすると

赤字：定値

$$Perplexty(\text{評価用コーパス}) = P_M(w_1 w_2 \ldots w_n)^{-\frac{1}{n}}$$

評価用コーパスを一定として複数のモデルを比較する場合，パープレキシティが低いモデルほど，評価用コーパスの尤度が高い（評価用コーパスを生成しやすい，もしくは次にくる単語を当てやすい）．

また，もう一つの視点

クロスエントロピー：

$$H(P_L, P_M) = -\lim_{n\to\infty} \frac{1}{n} \log P_M\,({w_1}^n)$$

近似

言語Lのエントロピー：

$$H(L) = \lim_{n\to\infty} -\frac{1}{n} \log P_L(w_1 w_2 \ldots w_n)$$

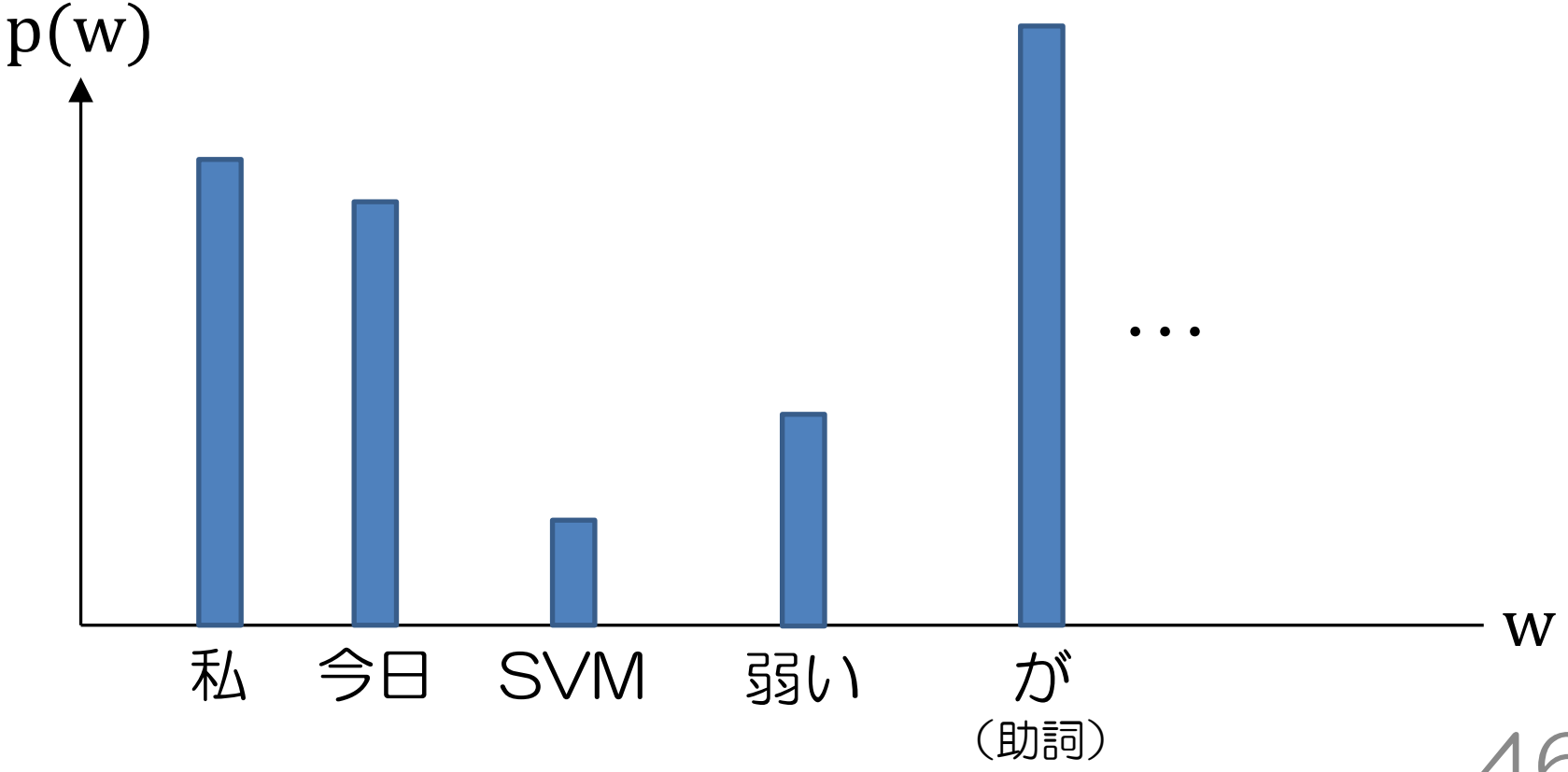
まとめ

- **言語Lのエントロピー：**
  - 言語L（情報源）から１個単語が出て来る際の不確定度.
- **パープレキシティ：**
  - 真の言語モデルまでの距離的な指標.
  - 小さいほど真の言語モデルに近いから良い.
  - 小さいほど評価用コーパスの尤度が高い.

もともと、数式というのはていねいに読まなくちゃいけないものなんだ。数式を読むときは、さらっと流してはだめ。じっくり読むことが大切。だから、そこに使われている文字をていねいにじっくり読むというのはいい態度なんだよ。

結城 浩：数学ガールの秘密ノート「式とグラフ」，第 1 章 文字と恒等式，p.4，ソフトバンククリエイティブ株式会社 (2013).

ユニグラム確率分布

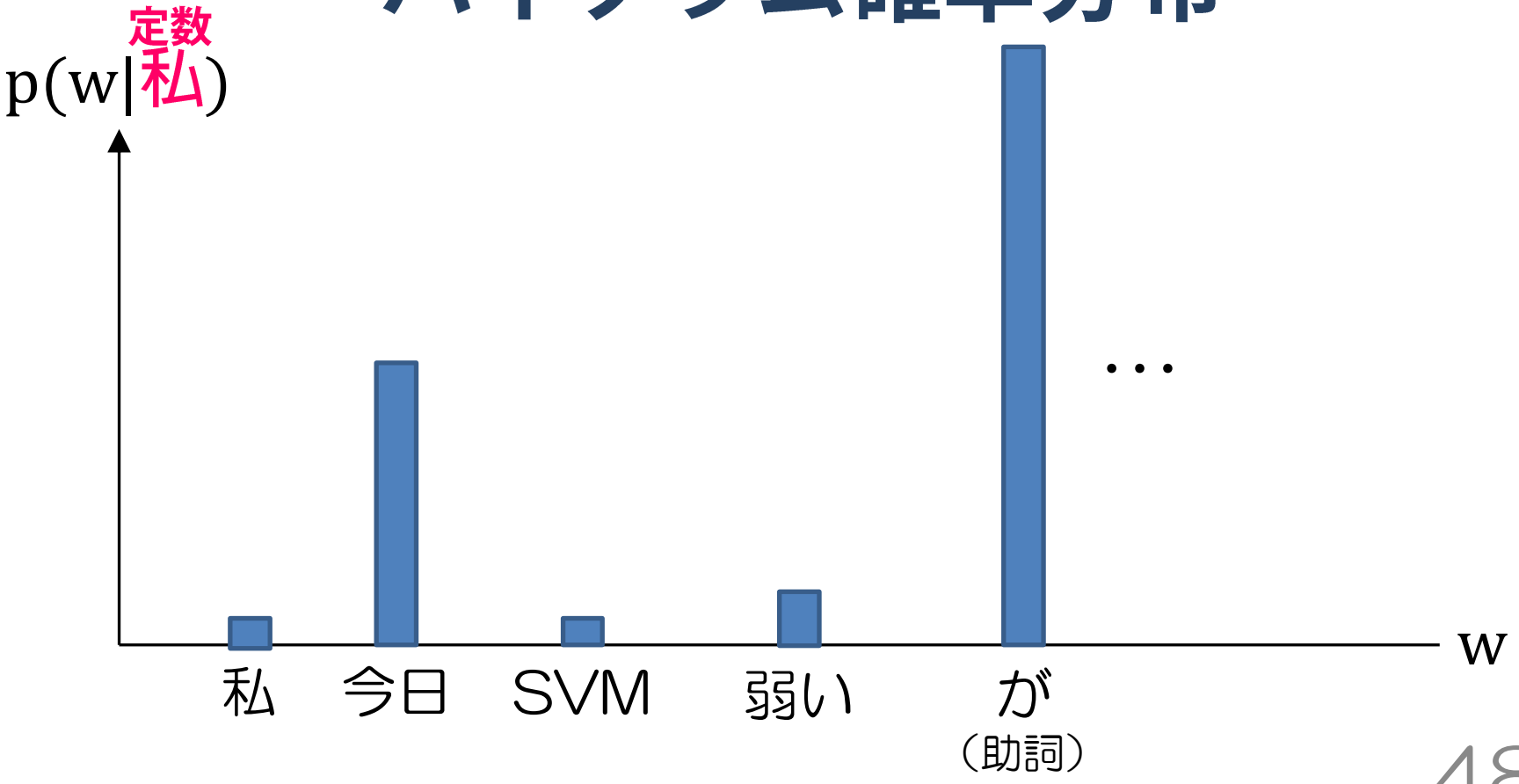


ユニグラム確率分布のエントロピー

- **ユニグラム確率分布p(w)のエントロピー：**
  - p(w)に従ってランダムに単語が１個もらえる状況で，「その単語が何か？」をどのくらい当てやすいか表す指標．

  - 「ユニグラム確率を基に計算したテストセット上のエントロピーではない」ことに注意！

バイグラム確率分布



バイグラム確率分布のエントロピー

- ある具体的な単語(e.g., 「私」)の次にどの単語が続くかをどのくらい当てやすいかの指標.

$$-\sum p(w|\text{私})\ log_2\ p(w|\text{私})$$

定数　　定数

- 「バイグラム確率を基に計算したテストセット上のエントロピーではない」ことに注意！