3. 冗長性と誤り検出・誤り訂正

冗長度

- 最大エントロピー H_{max} と実際のエントロピー H の相対的な差を表現するための指標として、冗長度 R を以下のように定義する

$$R = 1 - \frac{H}{H_{max}}$$

- 情報表現の効率の悪さを表現する指標として有用
  - 1に近いほど冗長であり、0に近いほど無駄がないことを示す
  - イメージ: 官公庁の分厚い文書の冗長度 ＞ ニュースの見出しの冗長度

自然言語の冗長性

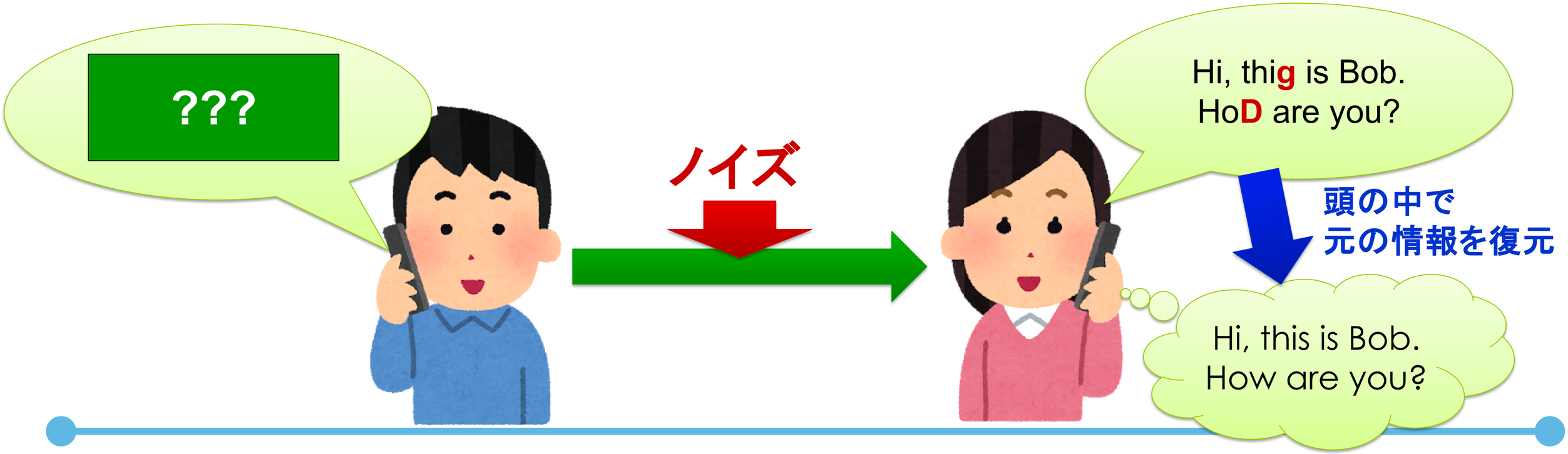
- 実は、日本語や英語などの自然言語は情報表現としてはかなり冗長
- 英語を例に挙げると
  - 文字の出現頻度に明らかな偏りがある
    - “E” は8文字に1文字を超える高確率で出てくるが、“Q” や “Z” は1000文字に1文字未満しか現れない
  - 文字 “Q” の後に来るのはほとんどの場合 “U”
    - 例: quarter, quadruple, question, quake, …
    - どうせ “q” の後は “u” なら、”qu” という一つの文字にしたほうが少ない文字数で済む

Letter	Relative frequency in the English language
e	12.702%
t	9.056%
a	8.167%
o	7.507%
i	6.966%
n	6.749%
s	6.327%
h	6.094%
r	5.987%
d	4.253%
l	4.025%
c	2.782%
u	2.758%
m	2.406%
w	2.361%
f	2.228%
g	2.015%
y	1.974%
p	1.929%
b	1.492%
v	0.978%
k	0.772%
j	0.153%
x	0.150%
q	0.095%
z	0.074%

図の出典: Case Western Reserve University
https://case.edu/artsci/math/singer/Sage%204/alphabet-frequency.html

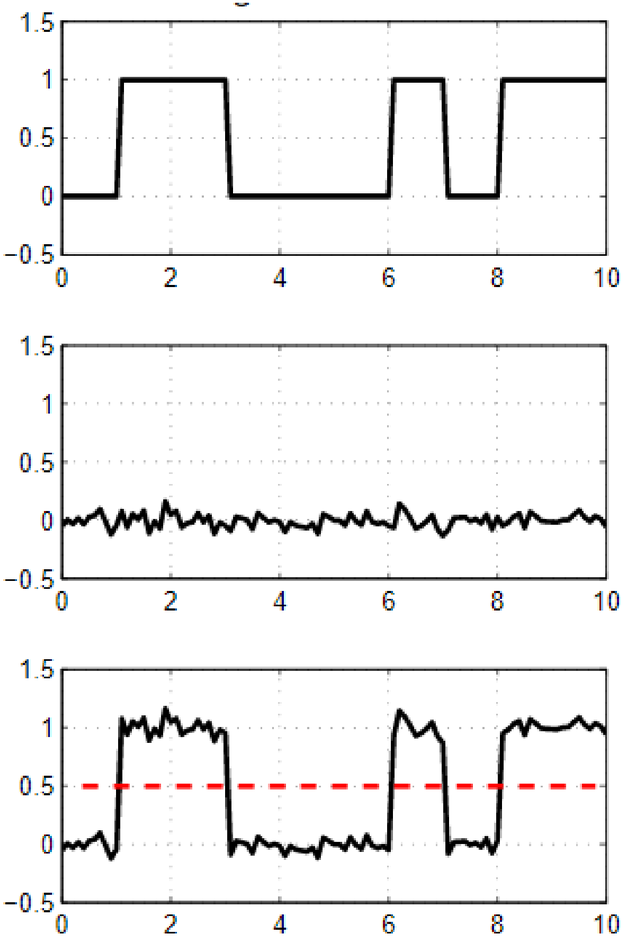
冗長性は単なるムダ？

- 実は、エラーやノイズによる情報の欠損を想定しなければいけない状況では、冗長性は有益でもある
- 通信路にノイズがある場合でも、通信されるデータに冗長性があれば、元の情報を復元できる可能性がある
  - 英語が冗長性を持つから、文字が欠損しても言いたいことが推測できる



データに意図的に冗長性を持たせる

- 一般に、無線をはじめとした通信においては様々なノイズが入る可能性がある
- 意図的にデータに冗長性を持たせることで、以下の2つのことが可能となる
  - 誤り検出
  - 誤り訂正



図の出典: Wikimedia

誤り検出符号

- 情報が一部壊れて受信または読みだされたときに、データの破損を検出できるような符号化を「誤り検出符号」という
- 例: パリティビット
  - オリジナルのビット列に対して「パリティビット」と呼ばれる冗長なビットを付加
  - パリティビットには、オリジナルに含まれる “1” の個数が偶数か奇数かを表す1ビットのデータを含める
    - 0 : 偶数個 , 1 : 奇数個
  - このとき、データが1ビットだけ誤って読みだされたときに、いずれかのビットが化けて伝わったことを検出できる

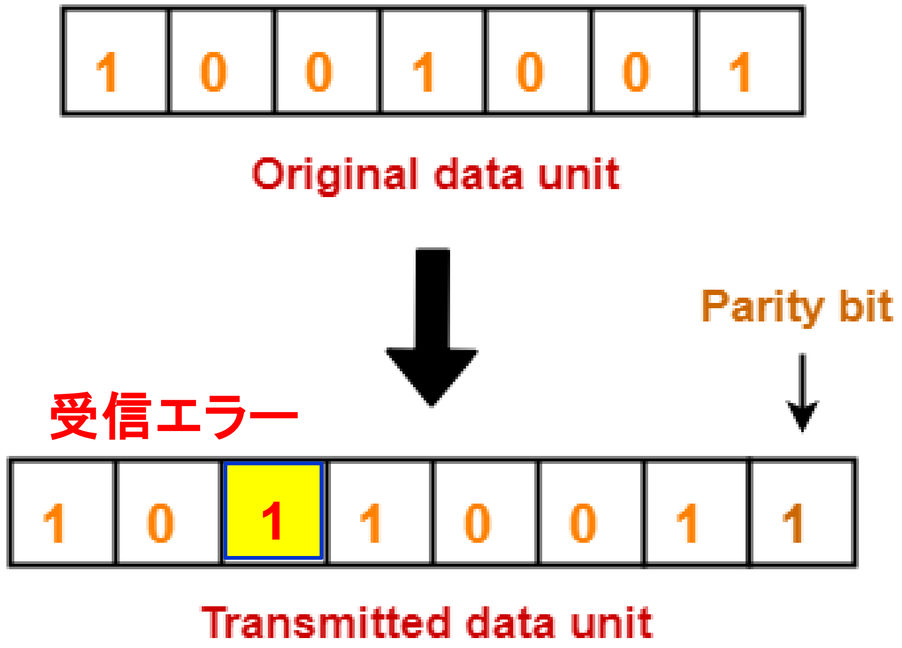
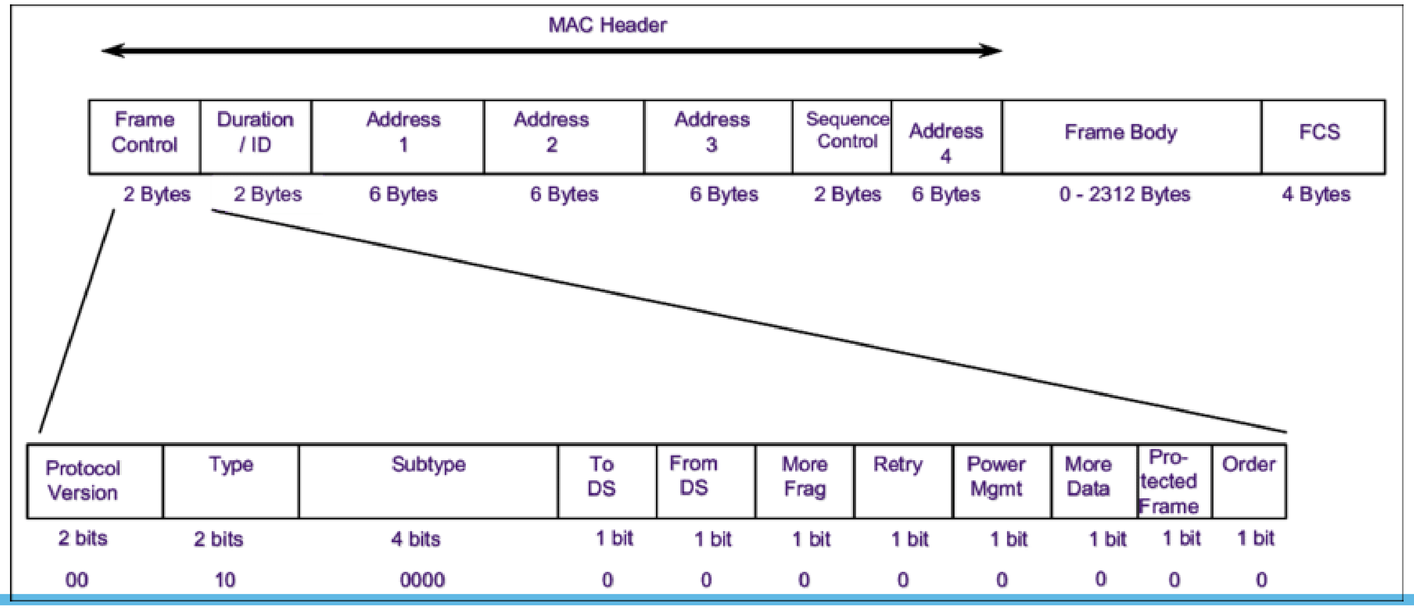


図: https://www.gatevidyalay.com/parity-check-parity-bit-error-detection/

誤り検出符号の利用例

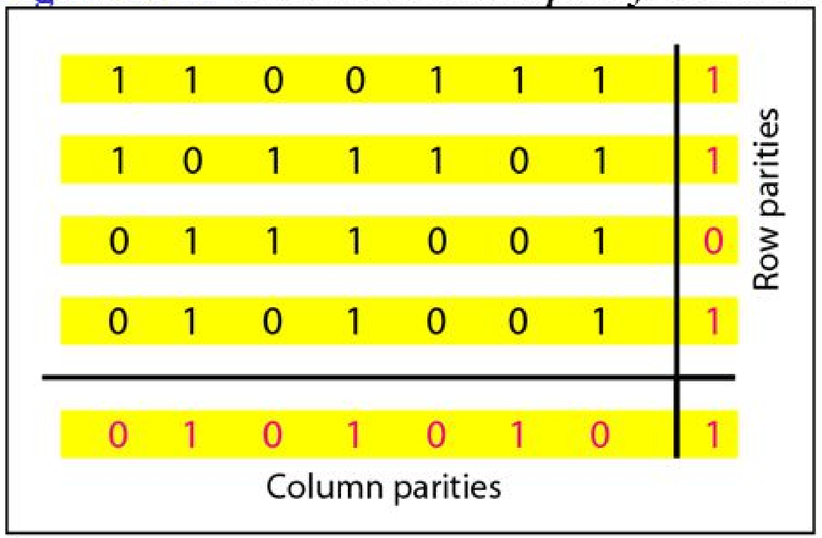
- 誤り検出符号は様々な通信に利用されている
- 例: Wi-Fi (IEEE 802.11) のフレームフォーマット
  - 4バイトの冗長な FCS (frame check sequence) をデータに付加することで、誤り検出を可能にしている
  - FCSによって誤りが検出された場合、そのフレームを破棄する仕組みとなっている

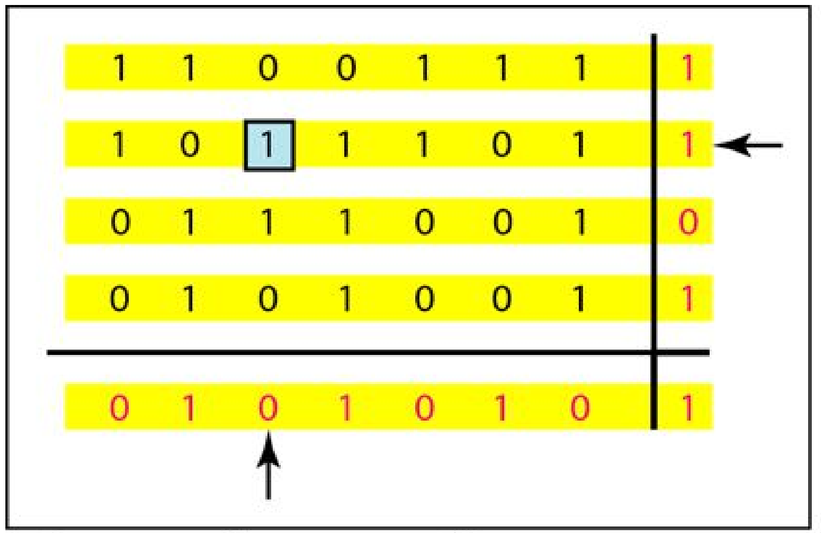


誤り訂正符号

- エラーの検出に加え、正しくオリジナルのデータを復元できるような符号化を「誤り訂正符号」という
- 例: 2次元パリティ
  - 二次元的に並べたデータの行と列両方についてパリティビットを持たせる方式
  - どの行・どの列でエラーが発生したかを検出できるので、そのビットを反転させれば元のデータを復元できる
  - 効率はとても悪いので、通常実用では利用されない

Figure 10.11 *Two-dimensional parity-check code*





b. One error affects two parities

図: Data Communications & Networking (B. Forouzan)

誤り訂正符号の例と応用

- ただし、2次元パリティは効率が悪いので、実用上は数学的な理論をもとにした符号が用いられている
  - BCH符号 (Bose-Chaudhuri-Hocquenghem符号)
  - RS符号 (Reed-Solomon符号)
    - 連続したビットの破損(バースト誤り)に対して強いのが特徴
- 誤り訂正符号は、様々な場面で利用されている
  - 地上波デジタル放送 (RS符号+畳み込み符号)
  - QRコード (RS符号)
  - ディスクデバイス (Blu-ray Disc, etc.)

ディスク・メモリのECC機能

- ハードディスクは衝撃に弱い
  - 読み書き中にちょっとした衝撃でヘッドの位置がずれてしまうと、間違った箇所にデータを読み書きしてしまうことに
- そのような場合でもデータが破損しないよう、通常ハードディスクにはECC機能が内蔵されている
  - ECC: error correcting code (誤り検出符号)
  - 皆さんのPCのハードディスクの中のビットはかなり化けているはずですが、正しいデータを読み書きできているのはECCのおかげです
- 同様に、ECCを搭載したRAMも存在する
  - 宇宙線などによるビット化けに対して耐性を持つ
  - 特に高信頼性が求められるシステムで用いられる

参考: シャノンの情報理論

- 本日扱った情報理論は、1940年代にクロード・シャノンによって確立された内容が中心となっている
- シャノンの理論は、データ圧縮や誤り訂正・誤り検出、暗号などについて議論する際の理論的な根拠をなしている
  - 本日扱っていない内容もたくさん含まれているので、興味のある人は自分で調べてみましょう