単文字換字暗号における暗文長・文字種類数と解読率との関係*

1S－5

大前 渉　乾 伸雄　野瀬 隆　小谷 善行　西村 恕彦
(東京農工大学 工学部 電子情報工学科)

1 はじめに

暗号解読率は、暗号文だけの攻撃法だと、暗号文の長さにだけ影響される。一般に、ある現象の頻度が n 倍になったとき、その分布が正規分布に従うとすると、中心極限定理より平均値の標準偏差 σ は、$\sigma / \sqrt{n}$ となることが知られている。

これを暗号解読に置き換えると、与えられる暗号文が n 倍になると、ある文字の頻度も n 倍になると予想される。すると、その使用率の標準偏差は、$\sigma / \sqrt{n}$ になるはずである。これが暗号解読率にどのように影響するのかを考察する。

2 文字種による特徴抽出

まず、文字種類数が変化することで、文字頻度特徴がどのように変わるのかを調査した。調査の対象にした文字種は、ローマ字・アルファベット・カタカナ・漢字かな混ざり文の4種類である。これらの文章データのうちカタカナ・漢字かな混ざり文は、別々に入手したもので、内容に関連性はない。ローマ字はカタカナの文章データを訓令式で直したものをもちいた。

図.1に4種の文字種の文章データの異なり語数の割合をしめす。これは、それぞれ文字数が100000字のときの異なり語数を100％としたときに、文字数によって、異なり語数がどのように変化するかを示したものである。ちなみに、100000字のときの異なり文字数は、ローマ字が20字、アルファベットが26字、カタカナが57字、漢字かな混ざり文が2197字である。

これを見ると、ローマ字・アルファベット・カタカナは、全文字種類数がほぼ同じで、異なり文字数もほぼ同じような傾きを示し、1000文字程度でほぼ全文字が出現する。漢字かな混ざり文は、1000文字でも10％程度しか出現しない。

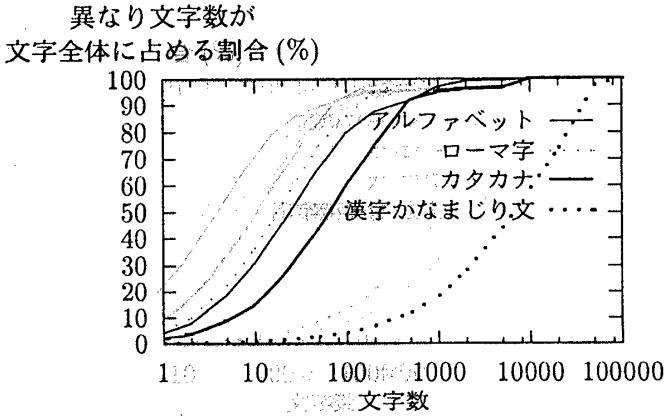

図.1:文字数と異なり文字数の関係

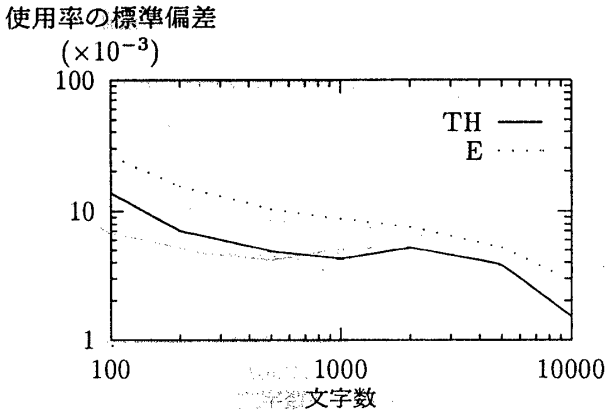

図.2:文字使用率の標準偏差の一例(アルファベット)

さらに、文字使用率の標準偏差を取った。その例を図.2に示す。ここでは、アルファベットの'E'と'TH'の文字数と使用率の標準偏差の関係を示してある。先ほど文字数が n 倍になると、ある文字の使用率の標準

*Success Rate of Cryptanalysis in Simple Substitution Cipher,
Wataru OHMAE, Nobuo INUI, Takashi NOSE,
Yoshiyuki KOTANI, Hirohiko NISIMURA,
Tokyo University of Agric. and Tech.,
Dept. of Computer Science

偏差は $\frac{\sigma}{\sqrt{n}}$ になるといったが、このグラフでもほぼそのとおりになっていることがわかる。これは、アルファベットだけではなく、カタカナなど他の文字種に対してもいえる。

3 解読実験

純粋に文字種類数と暗号解読率の関係を調べるために、解読プログラムには、ある言語特有の特徴を生かした暗号解読法を省いた。解読はアルファベットとローマ字を対象にして行なった。

解読法は[1]で提案された方法を改良したものである。

(1) 換字候補表を設定し、暗字側を暗文の単文字頻度順に並べる。母分布側には、なにも文字を設定しない。

(2) 次に、換字候補表の第 1 位から暗字に対応する平字を、最良優先探索で探索する。

(3) 探索の際の評価値には、暗字側と平字側の単文字・2 文字連接・3 文字連接の距離をあてる。母分布・暗文においての頻度順で i 番目の文字列の頻度を FP_i, FC_i とすると、先ほど述べた距離 D は次式で表せる。

$$D = \sqrt{\sum_{i=0}^{p} {\theta_i}^2 + \sum_{i=0}^{p}\sum_{j=0}^{p} {\theta_{ij}}^2 + \sum_{i=0}^{p}\sum_{j=0}^{p}\sum_{k=0}^{p} {\theta_{ijk}}^2}$$

(ただし、$\theta_i = FP_i - FC_i$)

また、暗文長と解読率の関係を調査するため、100 ～ 10000 字で構成される 10 種類の暗号文を複数用意し、解読させてみた。

文字長と解読率の関係を図.3 に示す。ここでいう解読率は、(解読に成功した文字種類数) / (暗文の文字種類数) を示す。

4 考察

4.1 暗文長と解読率の関係について

図.3 から、ローマ字の場合、暗文の文字数が 200 字までのときにと、200 字以上のときでは、解読率の傾きに違いが見られる。図.1 を見ると、200 字で、異なり文字数が文字全体の 95 %でほぼで揃う。アルファベットの場合も、2000 字までと 2000 字以上では、解読率の傾きに違いが見られ、異なり文字数も文字全体の 96 % でほぼで揃う。

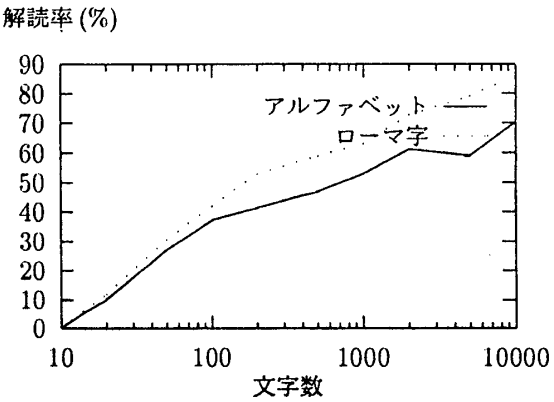

図.3:解読率

これより、異なり文字数が 95 %以上になると、解読率にも変化が見られるのではないかと考える。

4.2 文字種類数と解読率の関係について

今回は、ローマ字とアルファベットを対象として暗号解読実験を行なったわけであるが、これらは文字種類数にあまり差が見られず文字種類数と解読率の関係を求めることができなかった。今回の実験対象の文字種類数とは差がある漢字かな混ざり文・カタカナに対して解読実験を行なうことにより、文字種類数と解読率の関係を考察したい。

5 おわりに

本稿では、文字種類数および暗文長と暗号解読率の関係について、単文字換字暗号の解読実験を通して考察した。今後、漢字かな混ざり文・カタカナに対しても同様の実験を行う予定である。

参考文献

[1] 平塚 隆 : 換字式暗号の解読実験,
情報処理学会 第 22 回全国大会 講演論文集 3F-10 pp.19～20, 1981.

[2] 青木 利夫, 吉原 健一 : 統計学用論, 培風館, 1978.