日本語単文字換字暗号の解読

6J-3

下田あけみ，野瀬隆，西村恕彦
(東京農工大学 工学部 数理情報工学科)

1. はじめに

東京農工大学工学部数理情報工学科西村研究室では、1980年度から古典的暗号についての研究を報告している。今回は日本語（２バイトコードの文字）の単文字換字暗号の解読の方法について述べる。

2. 単文字換字暗号の原理

原文中の文字をある規則に従って他の文字や数字に変換する方式の暗号を、単文字換字暗号と呼ぶ。この規則は普通、原字と暗字を一字ずつ対応させた換字表で示される。ただし、日本語は文字の種類がJISコードに記載されているだけでも8836種ある。このため、まったくランダムな換字規則を用いると、8836対の原字－暗字の換字表を持たなければならない。そこで今回は換字規則として式①を採用した。式中のP（原字）、C（暗字）とは、各文字の内部表現コードである。暗号化の鍵はk_1、k_2、k_3である。この方式での暗号化の例を図1に示す。

$C \equiv k_1 * P + k_2 \pmod{k_3}$ …式①

ただし、 P：原字 C：暗字

k_1, k_2, k_3：鍵

7E7E(H) $< k_3 \leqq$ FFFF(H)

$1 \leqq k_1 < k_3$, $0 \leqq k_2 < k_3$

k_1とk_3は互いに素

原文　今 日 は 南 東 の 風 の 強 い 日 だ 。

↓ 数値化

14883, 18044, 9295, 18030, 17772, 9294, 18807, 9294, 13871, 9252, 18044, 9280, 8483

↓ 暗号化 (C ≡12*P+3 (mod 65521))

暗文　47557, 19968, 46022, 19800, 16704, 46010, 29124, 46010, 35413, 45506, 19968, 45842, 36278

図1 式①を用いた暗号化の例 （鍵 k_1=12, k_2=3, k_3=65521 の場合）

3. 暗号解読の原理

2.で示した換字規則から、鍵k_1、k_2、k_3を未知数とする3元連立方程式を解くことで解読ができる。この方程式を解くために必要なのは、原字と暗字の正しい組合せを3組見つけ出すことである。

4. 言語特徴

原字と暗字の正しい組合せを挙げるためには、日本語の統計的特徴を把握することが必要である。これを調べるにあたり、小説[1]と新聞記事[2]からそれぞれ10万字を得、単文字の頻度（表1）、2字列の特徴、文字種の含有率（表2）、ひらがなや漢字の性質などの統計調査を行った。

4.1 単文字頻度

表1に、単文字使用頻度のうち、頻度順で上位15位までのものを示す。

表1 単文字使用頻度（上位15）

a) 小説（総字数104,205字）

順位	文字	頻度(%)	順位	文字	頻度(%)	順位	文字	頻度(%)
1	、	5.02	5	っ	2.83	11	に	2.11
2	い	3.73	7	な	2.78	12	し	1.97
3	の	3.33	8	と	2.26	13	か	1.93
4	て	3.14	9	う	2.23	14	。	1.89
5	ん	2.83	10	た	2.15	15	る	1.61

b) 新聞（総字数100,000字）

順位	文字	頻度(%)	順位	文字	頻度(%)	順位	文字	頻度(%)
1	の	3.45	6	は	1.79	11	い	1.49
2	、	3.11	7	と	1.75	12	が	1.35
3	に	2.12	8	し	1.62	13	て	1.29
4	た	1.88	9	る	1.62	14	で	1.24
5	を	1.82	10	。	1.58	15	な	1.11

表2 各文字種の含有率

文字種	含有率(%)	
	小説	新聞
ひらがな	63.0	34.1
カタカナ	2.9	7.9
数字	0.02	4.7
英語大文字	0.01	0.6
英語小文字	0	0
漢字第一水準	20.6	43.6
漢字第二水準	0.06	0.04
特殊文字	13.5	9.0

Cryptanalysis of Japanese simple substitution cipher
Akemi SHIMODA, Takashi NOSE, Hirohiko NISIMURA
Tokyo University of Agriculture and Technology

4.2 候補文字の出現頻度

3.で述べたように、原字と暗字の正しい組合せが3組あれば、この方式の暗号は解読できる。4.1の調査から、正しい組合せのための候補文字を得た。ある文章について単文字頻度統計をとったとき、その上位には候補文字『、のいとたにてなしる。』が含まれている可能性が高い。そこで、候補文字がある文章の頻度統計上、第何位（又は、同順位もあるので、第何文字）までには、候補の中から3文字以上現れるかを調査した。

調査方法

① 小説、新聞それぞれについて、いろいろな長さ（1000字、500字、400字、300字、…）に区切った標本を用意する。各標本内に現れる単文字を出現度数順にソートする。

② 各標本内の順位において、上位何位までに、4.1で述べた6文字（又は11文字）中の3文字が現れるかを調べる。同じ長さの標本のうち80%以上が、上位n位までに候補文字3文字以上を含むようなnを求め、これを「候補順位数」とする。

③ 標本内の順位において、「上位n位」が上位何文字にあたるのかを調べる。例えば、度数1位の文字が同順位2個、3位の文字が同順位3個で、候補順位数nが4なら、4位以内の文字数は5、というように数える。同じ長さの標本について、この平均を調べる。こうして「候補順位数」に対応する文字数の平均値を「候補文字数」とする。

この標本の長さ、候補順位数、候補文字数を表3に示す。

表3 暗字側の候補文字数のめやす

a) 小説

標本の長さ(字)		1000	500	400	300	200	100	90	80	70	60	50	40
6字	候補順位数	7	7	7	7	8	9	9	9	8	11	11	9
	候補文字数	7.4	7.7	7.9	8.1	9.6	12.4	13.6	15.3	14.1	17.4	25.1	24.6
11字	候補順位数	4	4	5	5	5	5	5	5	5	5	5	7
	候補文字数	4.2	4.3	5.6	5.7	6.1	7.1	7.4	7.5	8.0	9.2	9.5	15.8

b) 新聞

標本の長さ(字)		1000	500	400	300	200	100	90	80	70	60	50	40
6字	候補順位数	5	6	6	6	7	10	11	13	12	12	10	10
	候補文字	5.4	6.9	7.0	7.4	9.3	17.7	17.1	19.7	21.3	41.1	36.7	33.8
11字	候補順位数	4	4	4	5	5	6	6	6	7	9	8	7
	候補文字数	4.3	4.6	4.7	6.0	6.6	10.5	11.5	11.8	12.3	25.6	28.6	29.1

『6字』とは『、のいとたに』との対応、『11字』とは『、のいとたにてなしる。』との対応

表3より、例えば小説の方のデータであれば、200字の暗号文を解読する場合、原字6文字に対し暗字の度数上位の10文字（あるいは原字11文字に対し暗字6文字）を対応させて組み合わせていけば、80%くらいの確率で正しい原字－暗字の組を3組得られると言えるだろう。対応させる原字の候補の文字数を6文字と11文字のどちらにするかは、暗文の文字数によって、計算量的に負担の少ない方に決定する。実際の解読システムでは、一般への適応を高めるために、表2の小説での値と新聞での値とで多い方の「候補文字数」を使用する。

5. 解読手順

4.の統計的特徴を用い、3.で述べた原理に基づいた解読の手順を次に示す。

① 暗号文に現れる単文字を度数順にソートする。度数上位の文字を暗字側候補とする。候補の文字数は表3を参考にして決定する。

② 4.1で述べた原字側の候補文字と、暗字側の候補文字を対応させて、原字－暗字の組を3組作る。

③ ②で対応させた3組の文字（数値）を式①に代入し、3元連立方程式を解き、未知数である鍵k_1、k_2、k_3を求める。

④ ③で求めた鍵を用いて暗文を100字程度復号し、平文らしさの検定を行う。平文らしくなければ、②に戻り、次の組合せで③、④を行う。

⑤ ④で平文らしいという結果がでれば、暗号文全文を復号し、数値から文字に変換する。

6. おわりに

今回述べた暗号方式による暗号文の解読の可能性は、3組の組合せの成功に依存している。そして、その組合せに利用した統計的特徴は主として単文字使用度数である。しかし、組合せの数からみて、数十文字の暗号文でも解読できる見通しを得ている。今回の調査・実験データは、日本語の、他の暗号方式の解読システムを考察する上で、基礎となるものである。

7. 資料及び参考文献

資料

[1] 新井素子：『結婚物語（上）』，角川書店，1986年 （第一章～第七章，pp.5～244）

[2] 情報処理学会第36回（昭和63年前期）全国大会発表（講演番号5B-10）『日本語の統計的特徴と日本語周期転置暗号の解読』で利用したアスキー社の時事新聞の記事データベースの一部

参考文献

[3] 松井甲子雄：コンピュータのための暗号組立法入門，森北出版，1986年

[4] 村松康弘：日本語暗号の解読法に関する研究，東京農工大学工学部数理情報工学科1987年度卒業論文