

第2回ナレッジグラフ推論チャレンジ 2019

小説「踊る人形」の暗号解読の検討

チーム名 AKK

メンバー:

富士通研究所 村上勝彦

神戸常盤大学 高松邦彦

神戸市立西神戸医療センター 杉浦あおい

- 応募カテゴリ: アイデア部門
- チーム名: AKK
- メンバー:
 - 富士通研究所 村上勝彦
 - 神戸常盤大学 高松邦彦
 - 神戸市立西神戸医療センター 杉浦あおい
- メールアドレス(代表): murakami.kt@fujitsu.com

- 本提案では物語「踊る人形」中に示される手書きの人形図形暗号の解読方法を検討した。
- 上記の暗号について、数理計画法を用いたアプローチによって、ホームズの推理の再現と説明を試みる。本提案では、暗号を機械的に解読する手順を検討して示した。
- ホームズの推理で曖昧な点に対し、数理的方法による解決法を盛り込んだ。
- 本アイデア全体は未実装だが、部分的なデータ作成や簡単な検討を実施した。
- 理論的には、ホームズの推定した答えが、本提案である程度高スコアの解として得られると予想するが、実際には未検証である。

- 本チャレンジ全体の目的は、説明付きで質問に答えるAIシステムの実現を目指すこと。
- 具体的なテーマは、ホームズ推理の再現と説明である。
- 解析対象の物語は5話提示されている。
- 本提案では5話全部ではなく、物語「踊る人形」1つに絞り、その中の手書きの図形暗号の解読に着目した。
 - チャレンジサイトで示された「個別タスク:踊る人形:暗号を解け(暗号の解読)」に相当。
- 上記の暗号について、人工知能的アプローチによって、ホームズの推理の再現と説明を試みる。

基本的な問題設定と概要

- 物語「踊る人形」では、下図の6個のメッセージが順番に提示される。
- 人形の下に書かれたアルファベットは正解を示す。人形1つと文字1つが、図のように対応することが物語後半で解明される。
- ホームズは5個目までである程度解読し、最後(6個目)のパターン (come here at once) を作成したので、5個から解読できるのが望ましい。

①		②		③	
	AM HERE ABE SLANEY		AT ELRIGES		COME ELSIE
④	NEVER	ELSIE PREPARE TO MEET THY GOD ⑤			
⑥					
	COME HERE AT ONCE				

http://www.chikyukotobamura.org/muse/wr_fiction_19.html

ホームズの暗号解読ポイント

■ 前提: 記号は文字の代用とする

- (Homesは暗号の類型を熟知、小説書いたり、160種の暗号法分析の経験)

理由まとめ

■ 文字頻度による推理: 2回 (客観性 高)

- アルファベットでは E がもっともよく使われる
- 次に使われるのは T, A, O, I である

■ 被害者名、3度末尾に登場: 1回 (客観性 高)

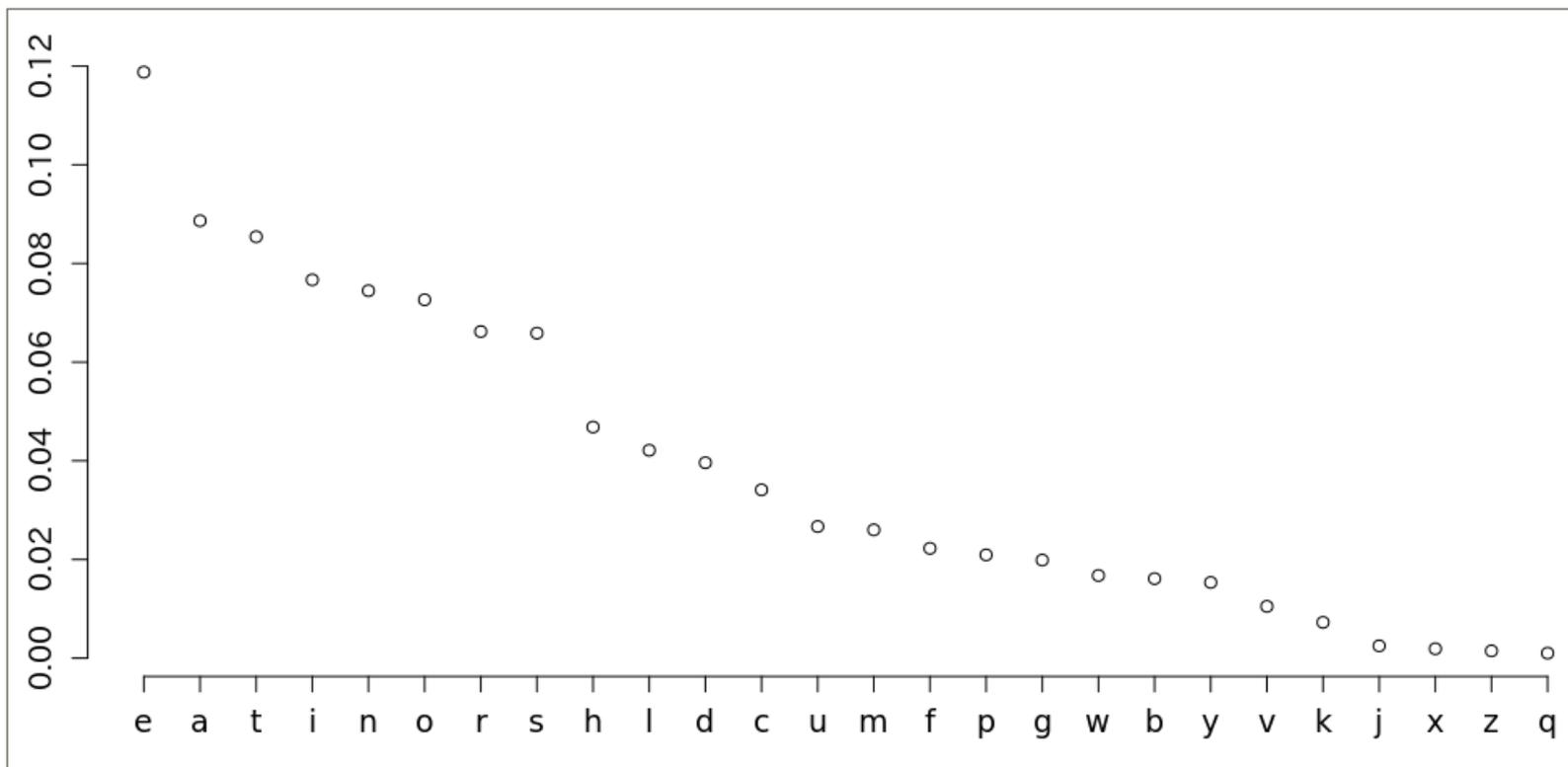
■ 返事としてありそう、意味が通じる、状況に合う: 6回 (客観性 低)

■ 以下の疑問をまず検証

- 文字頻度の情報は推理の通りに本当に利用できるか?
- 人形図のパターンで、同じ文字かどうかどうやってわかるか不明

アルファベットの出現確率が使えるか

(Wikipedia text 11,055,558,613文字から独自集計)

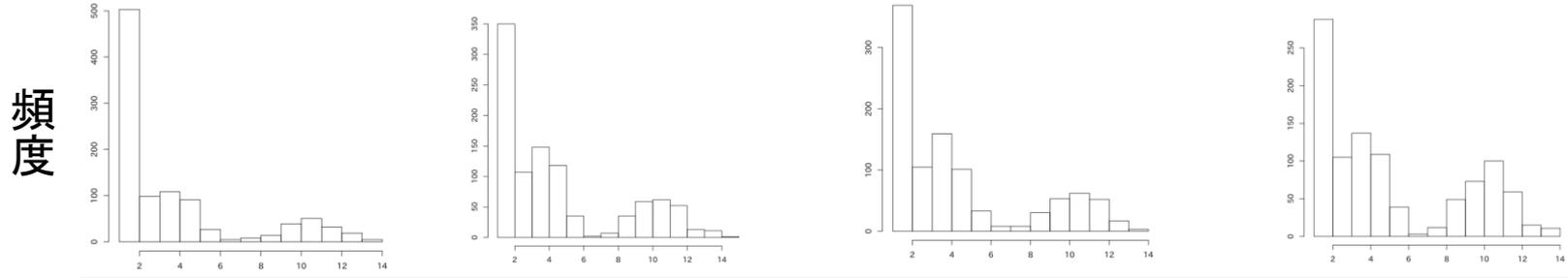


- 最高頻出文字は”e”で、小説通り1位で突出している。
- 次に、”a”, “t”の2つ、または7つが近い頻度で徐々に次第に下がる(e, a, t, i, n, o, r, s の7個が高い)

トップ4文字(e,a,t,l) が頻度の順で何位になるかシミュレーション

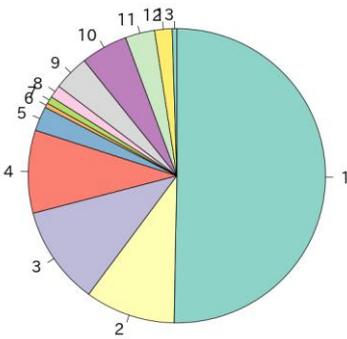
- 1: 順位分布をプロット
- 2: 円グラフで割合を検証した
- シミュレーション: 先の分布で多項分布、回数は1000回

()

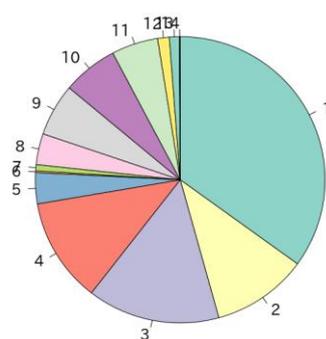


順位(1~15)

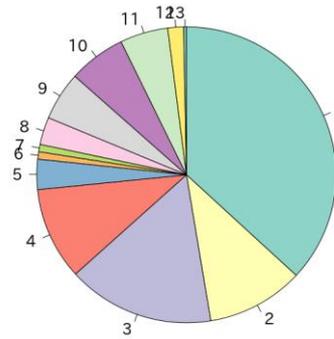
数字は
順位



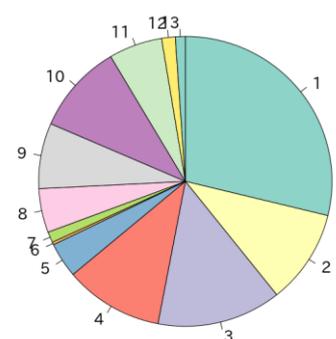
E



A



T



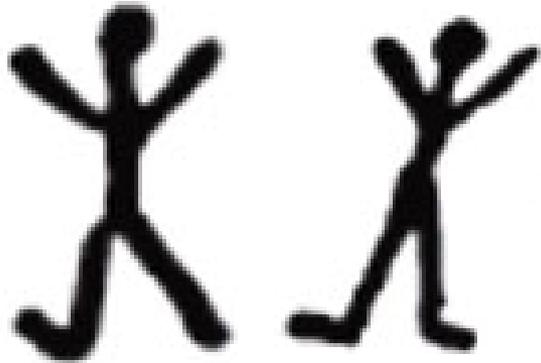
I

Eが1番多く出現するのは半分のケースしかない(左図)。A,Tも一位になるときが多い(同率含む)。「1番多い文字がEであろう」という方法(仮定)は半分正しいが、一般的にはそのまま使えない。

画像解析問題の難しさ: これらを区別すべきかどうか

「これらを区別すべきか」は、「特徴変数を付与すべきか」と言い替えてもいい

例1



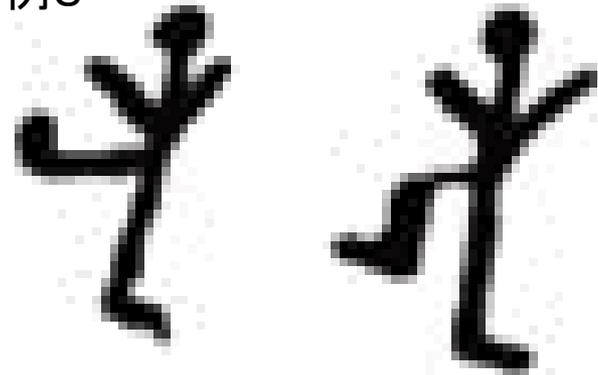
足はね、体の傾きが同じか判断が難しい。
小説では、E, Eで同じとされている。

例2



足が浮いていて同じか、膝をまげが違うのか、
かどうかの判断が難しい。
小説では、Aで同じと解読されている (a-1-1,
a-1-12)

例3



足が浮いていて同じか、膝をまげが違うの
か、判断が難しい。
小説では、I, Oと違うとされている (I-5-4,
O-5-23)

- 物語中では解説がなかったが、以下は謎である
 - どうして似た文字が同じ文字だとわかるか？
 - 区別すべき人形図の特徴は事前に分からない。
(言い替えると、似ている人形が同じ文字かは自明でない)
 - 図の特徴には、文字種に関係ないものが含まれる(例:旗、体の傾き)
 - 26文字すべては表れていない
- そこで上記謎も含めて、以下の問題を解けるのがよい
 - 人形図とアルファベットとの対応付け →文章が決定される
 - 人形図を判別するための部分特徴が何か

■ 方法

1. **人形図の部分的特徴をコード化する(ここは人の判断)**
 1. 上記特徴を使用するかどうかを二値変数で設定
 2. 図と文字の対応を決定すべき二値変数
2. 比較的**小さい辞書(候補となる単語リスト)**を作成
 1. 動詞は原型のみ、文字数は2から8程度
3. 以下の状況にあいそうな単語に**より高いスコア**を付与する
 - 慣用句
 - 文字数が多い単語
 - 要求時のメッセージに出てきそうな単語に重み付けしておく
 - 物語の固有の人名、場所名
 - 金銭的価値を示す単語(財産、銀行、取引、犯行計画、金庫等)
 - 感情(愛情、怨念等)を伝える言葉
 - 動詞(命令、要求を伝えるため)
 - 悪口、危害(おびえる反応を引き起こしている)
4. 文字列あてはめ位置と特徴量を変数とし、**最適化**により最適解を求める
制約条件として「**同じ文字は暗号が一定以上近い**」とする

暗号同士の類似度コード化

- 以下では人間の認識する範囲で、人形図の特徴をデータ化した。
- 同じ文字同士は、同じ人形の特徴を持つ(右足上げる、など)
- 他の文字動詞は、違う特徴セットでないといけない
- 本文では、「旗の意味は単語の区切り」であるが、**導出可能**
- 特徴を最適化の変数(1, 0)として利用するため、使わない特徴があっても無視するか判定可能**

正解文字	出現位置													
	右腕		左腕			右足		左足		体の向き		旗		
	有・無	曲・伸 (有の場合)	上・下 (曲の場合)	有・無	曲・伸 (有の場合)	上・下 (曲の場合)	股関節 曲・伸	膝曲・伸	股関節 曲・伸	膝曲・伸	上・下	有・無	右・左	
A	1_1	1	-1		1	-1		-1	1	-1	-1	1	-1	
A	1_7	1	-1		1	-1		1	1	-1	-1	1	-1	
A	1_12	1	-1		1	1	-1	1	1	-1	-1	1	-1	
A	2_1	1	-1		1	-1		-1	1	-1	-1	1	-1	
A	4_9	1	-1		1	-1		-1	1	-1	-1	1	-1	
A	5_10	1	-1		1	-1		1	1	-1	-1	1	-1	
B	1_8	-1			-1			1	1	-1	1	1	-1	
C	3_1	-1			1	-1		-1	1	1	1	1	-1	
C	4_1	-1			1	-1		-1	1	1	1	1	-1	
C	4_13	-1			1	-1		-1	1	-1	1	1	-1	
D	5_24	1	-1		-1			-1	-1	-1	-1	-1	-1	
E	1_4	1	-1		1	-1		-1	-1	-1	-1	1	-1	
E	1_14	1	-1		1	-1		-1	-1	-1	-1	1	-1	
E	2_3	1	-1		1	-1		-1	-1	-1	-1	1	-1	
E	2_8	1	-1		1	-1		-1	-1	-1	-1	1	-1	
E	3_5	1	-1		1	-1		-1	-1	-1	-1	1	-1	
E	3_9	1	-1		1	-1		-1	-1	-1	-1	1	-1	

データ化したもの

... 以下、76文字分続く(黄色は文字Aと文字Eの塊)

■ 目的関数(最大化)

$$f(x) = A + B$$

A : 単語のスコア

B : 同じ文字とした図同士の類似度

■ 変数

- $m_{i,j}$: 文字 i , 図 j が一致するなら $m_{i,j} = 1$, $\sum_j^n m_{i,j} = 1$.
- $u_{k,l}$: 文字 i , 図 l が一致するなら1、それ以外は0.
- $s_{o,p}$: 特徴 o が図 p にみられるなら1、それ以外は0。(類似コード)
- $w_{q,r}$: 単語 q が図 j から始まるなら1、それ以外は0.

- ホームズの推理はさまざまな人間的洞察(雑学知識)を用いるが、本提案では計算処理(辞書探索とスコア計算)や数理手順に基づいているため、説明性は高いだろう
- 正解を含めた小さい辞書で実行すれば現実時間で「正解」が出ると考えられるが、語彙を増やしていけば、「正解」以外の解が得られるかもしれない。
- 本方法の解析によって、以下のことが明らかになる可能性がある
 - 客観的に提示された情報だけでは解読は困難
 - ・ 提示された情報の条件が曖昧
 - ・ 解読を大きく助ける条件・情報は、解決と同時か事後に示されていた
- 本方法では、文脈に沿った解にするのは十分考慮してない

- 本提案では、暗号を機械的に解読する手順の概要を示した。
- 本提案において、暗号問題を解く数理的な手順の部分は、既存の最適化手法である。
- 小説でホームズが解いた暗号解読問題を、定義して数理問題にマップしたところ。
- 特徴
 - 辞書を小さくしぼることで計算量をしぼった
 - ありそうな単語のスコアで文脈考慮をいれた。
 - 「人形図を判別するための部分特徴が何か」にも回答できる
- 理論的には、ホームズの推定した答えが、ある程度高スコアの解として得られるはずである。
- 現時点では、未実装でありどの程度の時間で正解がでるかは、未検証である。引き続き取り組み、実験につなげたい。



FUJITSU

shaping tomorrow with you