Locally Typical Sampling

Clara Meister, Tiago Pimentel, Gian Wiher, Ryan Cotterell
TACL 2023

読む人: 持橋大地
統計数理研究所
最先端NLP 2023
2023-8-27

どんな話？

- Locally Typical Sampling:
  Nucleus samplingやTop-*k* sampling よりも高性能で理論的背景のある、
  言語モデルからの新しいサンプリング法
- 情報理論での「典型集合」(Typical set) の考え方を1単語ごとに局所的に適用
  - 実際の言語データからの認知的なエビデンス
- 言語モデルからの生成の際に、無意味な繰り返しや短すぎる文を避けて、より高精度に生成できる

背景

- 言語モデルからの生成は、大規模言語モデルの使用に不可欠の要素
- 確率のもっとも高い単語列を選ぶと、空文字列になったり (Stahlberg+19)、情報量の少ない文になる
- 単純にランダムサンプリングをすると、たまたま確率の低い語が生成された後が壊滅
  → Top-*k* sampling (ACL 2018)
  → Nucleus sampling (ICLR 2020) [最先端NLP2020]
- しかし、これらでもまだ無意味な繰り返しなどが生成されてしまう

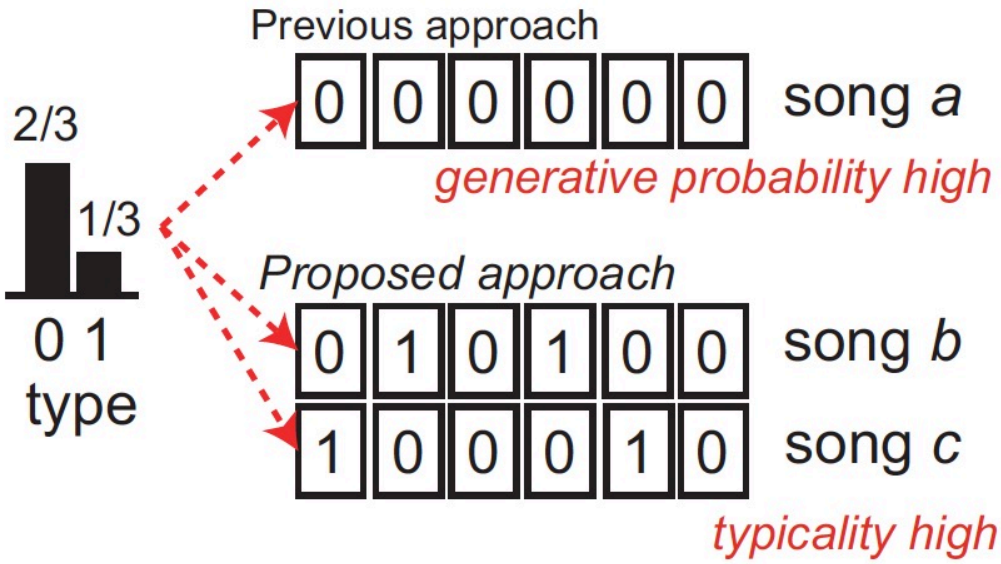
情報理論と典型系列

MUSICAL TYPICALITY: HOW MANY SIMILAR SONGS EXIST?

Tomoyasu Nakano[1] **Daichi Mochihashi**[2] **Kazuyoshi Yoshii**[3] **Masataka Goto**[1]
[1] National Institute of Advanced Industrial Science and Technology (AIST), Japan
[2] The Institute of Statistical Mathematics, Japan
[3] Kyoto University, Japan
[1] {t.nakano, m.goto}[at]aist.go.jp [2] daichi[at]ism.ac.jp
[3] yoshii[at]kuis.kyoto-u.ac.jp

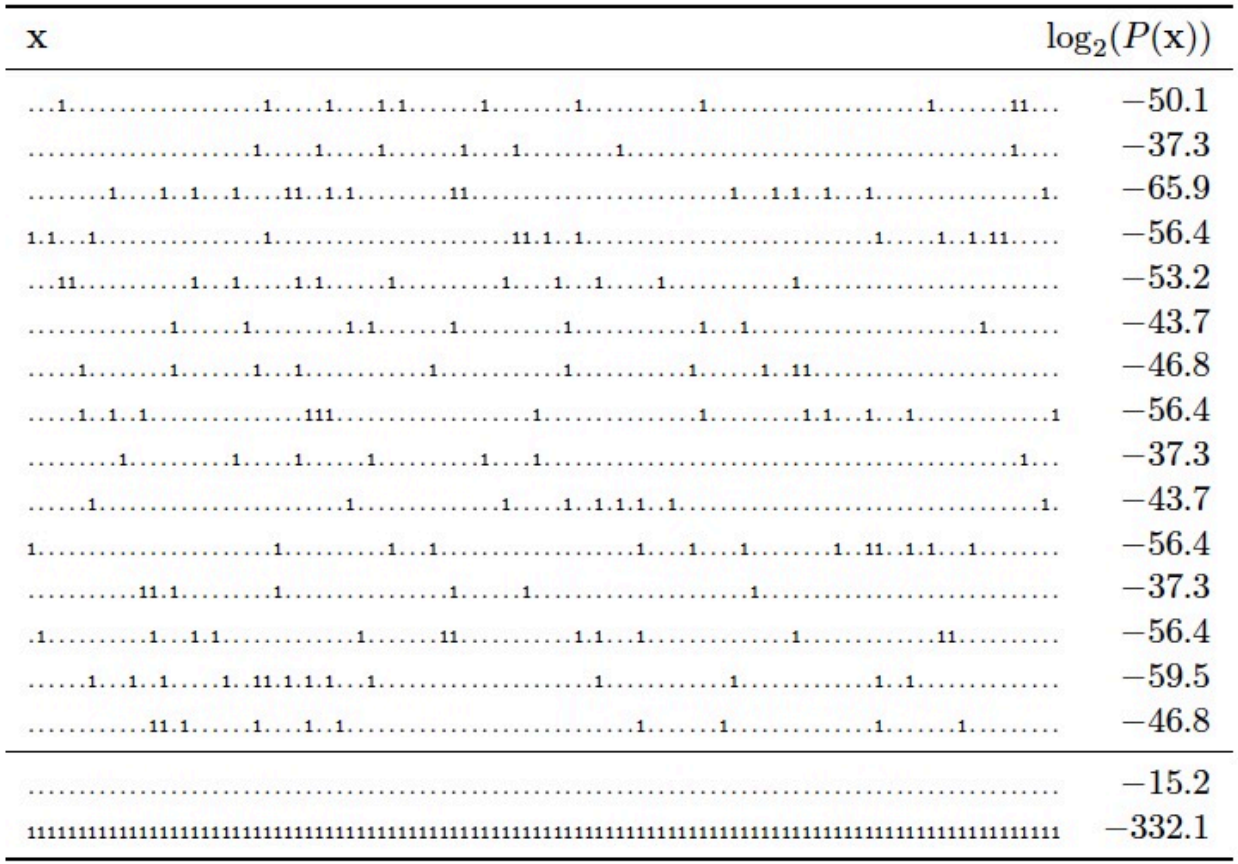
- ISMIR 2016 (音楽情報処理のトップ国際会議)での共著論文
  - 産総研音楽グループの中野さんとの共同研究

情報理論と典型系列 (2)



- 確率2/3で0, 1/3で1を出す情報源があったとして、
  そこから出る典型的な系列は0と1が混じっているはず
- 「もっとも確率の高い」0000...00 の確率は、
  指数的に小さい→実質的に、ほとんど出力されない

情報理論と典型系列 (3)



x	$\log_2(P(\mathrm{x}))$
...1.....1.....1....1.1.....1.....1.....1.....1.....11...	−50.1
.....1.....1.....1.....1.....1.....1.....1.....1....	−37.3
.....1....1..1...1....11..1.1.....11.....1...1.1..1...1.....1.	−65.9
1.1...1.....1.....11.1..1.....1.....1.....1..1.11.....	−56.4
...11.....1...1.....1.1.....1.....1....1..1.....1.....1.....	−53.2
.....1.....1.....1.1.....1.....1.....1...1.....1.....	−43.7
.....1.....1.....1...1.....1.....1.....1.....1..11.....	−46.8
.....1..1..1.....111.....1.....1.....1.1...1...1.....1	−56.4
.....1.....1.....1.....1.....1....1.....1...	−37.3
.....1.....1.....1.....1..1.1.1..1.....1.	−43.7
1.....1.....1...1.....1....1....1.....1..11..1.1...1.....	−56.4
.....11.1.....1.....1.....1.....1.....1.....	−37.3
.1.....1...1.1.....1.....11.....1.1...1.....1.....11.....	−56.4
.....1...1..1.....1..11.1.1.1...1.....1.....1.....1..1.....	−59.5
.....11.1.....1....1..1.....1.....1.....1.....1.....	−46.8
.....	−15.2
11111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111	−332.1

Figure 4.10. The top 15 strings are samples from X^{100}, where $p_1 = 0.1$ and $p_0 = 0.9$. The bottom two are the most and least probable strings in this ensemble. The final column shows the log-probabilities of the random strings, which may be compared with the entropy $H(X^{100}) = 46.9$ bits.

- MacKay “Information Theory, Inference, and Learning Algorithm” p.78: 1と0を(0.1,0.9)で出力する情報源

典型系列と典型集合

- 言語モデルに含まれるほとんどの文の確率は、エントロピーが一定
  → 典型系列の集合＝典型集合
- 確率最大の文が典型集合に含まれる確率は、ほぼゼロ
- エントロピーとは？

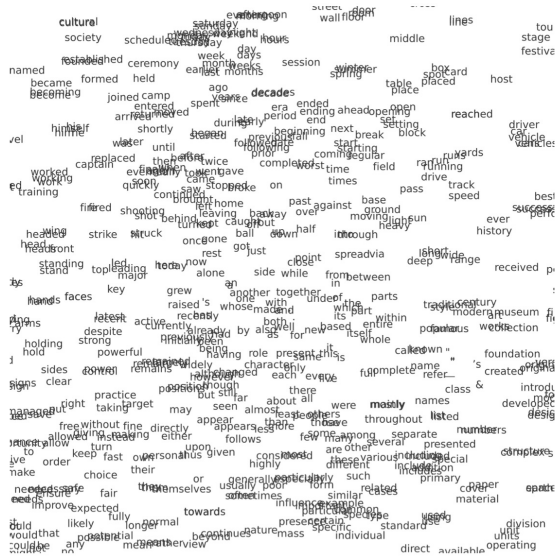
情報理論と自然言語処理

統計的テキストモデル

持橋 大地
統計数理研究所 数理・推論研究系
daichi@ism.ac.jp

2023 年 8 月 25 日

$Id: textmodel.tex,v 1.16 2023/04/04 16:27:30 daichi Exp $



2.6.3　情報理論の基礎

式(2.58)で用いた, 確率の対数

$$\log p(c_t|c_1,\cdots,c_{t-1}) \tag{2.62}$$

は, 単に計算の利便性だけでなく, 情報理論において**情報量**という意味を持っています. このことについて, 少し説明しましょう.

たとえば, 可能な選択肢が 8 つあってどれも等価なとき, どれかが選ばれる確率は 1/8 です. つまり逆にいうと, 確率 $1/8 = 0.125$ とは, 等価な選択肢が 8 つある状況と同じであることを意味しているわけです. この等価な選択肢の数のことを, **分岐数**といいます.

このとき実際に選択肢から 1 つを選ぶのに, 8 回の選択をする必要はありません. 8 個の選択肢を $x_1, x_2, \cdots, x_8$ と表すと, コインを投げて表裏のどちらが出たかを使うことにして,

- 1 回目のコインで $(x_1,\cdots,x_4)$ と $(x_5,\cdots,x_8)$ のどちらを選ぶかを決め, 前者の $(x_1,\cdots,x_4)$ が選ばれたとすると
- 2 回目のコインで (x_1,x_2) か (x_3,x_4) のどちらを選ぶかを決め, 後者の (x_3,x_4) が選ばれたとすると
- 3 回目のコインで x_3 と x_4 のどちらを選ぶかを決め, ここでは表が出たので, 最終的に x_3 を選ぶ

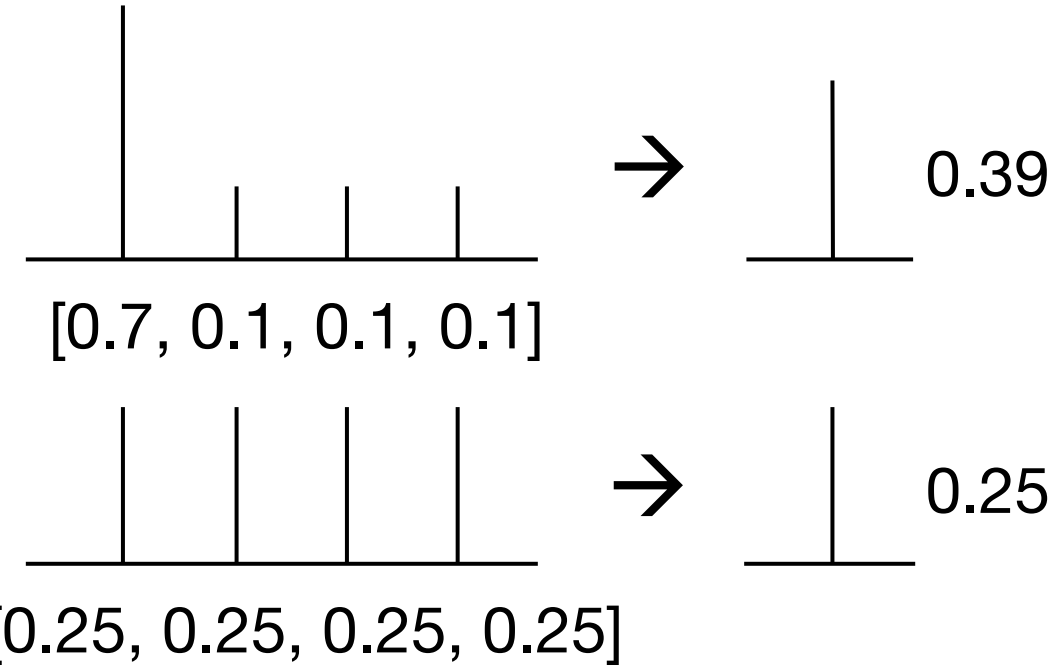
のようにすれば, どれかをランダムに選ぶことができます. このように, 8 個の選択肢であれば $\log_2 8 = 3$ 回のコインを振れば出力を決めることができるため, 必要な情報量は 3 だということができます. 同様にして, 確率 p の事象は $\dfrac{1}{p}$ 個の等価な選択肢から 1 つを選ぶことと等価で, このときの手数, すなわち必要な情報量は $\log \dfrac{1}{p}$ です. そこで, 確率 p の事象の「**情報量**」を表すこの量を Shannon の**自己情報量** (self information) といい,

*1　同時確率とはみなせない, 複数のモデルによる確率のような場合は, 算術平均をとっても問題ありません.

エントロピー＝(対数)平均確率

$$H(x) = -\sum_x p(x) \log p(x) = \left\langle -\log p(x) \right\rangle_{p(x)}$$

- よって、$\exp(-H(x))$ は「平均的な確率」を表す



典型集合

- ある情報源から出る長さTの系列で、対数確率がその期待値とε以内であるような系列の全体

$$\mathcal{T}_\epsilon^{(T)} = \left\{ \mathbf{y} \,|\, \left| \frac{-\log p(\mathbf{y})}{T} - H(\mathbf{Y}) \right| < \epsilon \right\}$$

- Tが大きくなると、
  - 典型集合に含まれない系列が出る確率は0に収束する
  - 典型集合内の系列の確率はどれも、ほとんど期待値と同じ (漸近等分割性)

局所典型系列

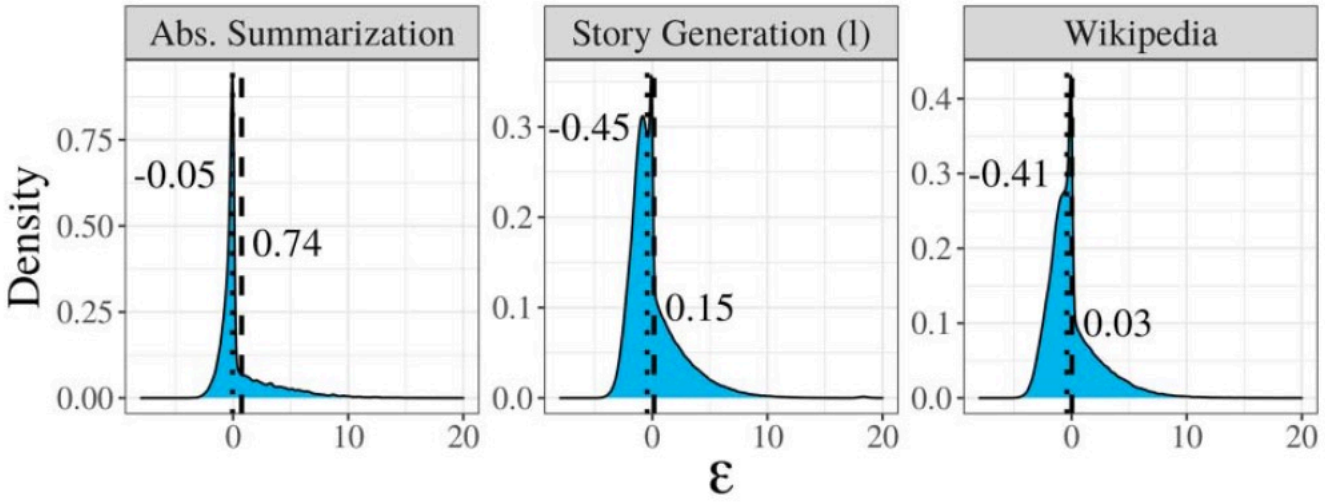
- 典型系列/典型集合は、1つの系列全体の確率を考えていた
  - 普通に言語モデルからサンプリングすると、これらの系列が得られる
- 実際の人間の言語は、文末があるのでエルゴード的ではなく、情報の無駄をなくすために
  各単語の条件付き確率が、その期待値に近い
  と考えられる
  - これを局所典型性 (local typicality)と定義する

実際の観測データ

- 本論文のタスクで、言語モデルで計算した次の単語の確率の対数と、予測分布からの期待値との差

$$-\log p(y_t|\mathbf{y}_{<t}) - H(Y_t|\mathbf{y}_{<t})$$

の分布



- 左の数字=中央値、右の数字＝平均
  どちらも、ほとんど0

アルゴリズム

- 各時刻tで、予測分布のエントロピー $H(Y_t|\mathbf{y}_{<t})$ を計算する
- 予測確率の対数 $p(y_t|\mathbf{y}_{<t})$ との差の絶対値が小さい順に、単語を予測集合に追加
- 確率の総和がτになったら終了
- 予測集合の中から、$p(y_t|\mathbf{y}_{<t})$ に従って単語をランダムにサンプリング

アルゴリズム' (確率版)

- 各時刻tで、予測分布から予測確率の期待値 $\bar{p} = \exp(-H(Y_t|\mathbf{y}_{<t}))$ を計算
- 各単語について、予測確率 $p(y_t|\mathbf{y}_{<t})$ と期待値 $\bar{p}$ との比

$$\frac{p(y_t|\mathbf{y}_{<t})}{\bar{p}}$$

が1に近い順に、単語を予測集合に追加

- 注意：
  - Nucleus sampling, Top-*k* samplingでは、単純に確率の大きい方から単語を予測集合に追加する

実験

- 文生成と要約の両タスクで実験
- Nucleus sampling, Top-*k* sampling, 温度付きサンプリング, Mirostatと比較
- 評価指標：
  - REP (Welleck+ 2020)：繰り返しが多いかのスコア
  - D：*n*グラムの多様性 (*n*=1..4)
  - Zipf：Zipf係数、ロングテールの度合い

文生成タスク

	Story Generation						
	PPL (g)	PPL (i)	MAUVE (↑)	REP (↓)	Zipf	D (↑)	Human (↑)
Reference	16.33	26.71	–	0.28	1.09	0.85	4.12(±0.02)
Temperature (τ=0.5)	25.34(+9.01)	18.78(−7.93)	0.95	**0.25**	**1.07**(−0.02)	**0.87**	4.13(±0.02)
Temperature (τ=1)	25.67(+9.34)	11.77(−14.94)	0.95	0.26	**1.07**(−0.02)	**0.87**	4.13(±0.02)
Nucleus (η=0.9)	7.75(−8.58)	10.25(−16.46)	0.95	0.35	1.29(+0.20)	0.79	4.09(±0.02)
Nucleus (η=0.95)	11.65(−4.68)	11.77(−14.94)	0.95	0.30	1.20(+0.11)	0.84	4.13(±0.02)
Top-k (k=30)	7.07(−9.26)	18.78(−7.93)	0.88	0.35	1.41(+0.32)	0.80	4.13(±0.02)
Top-k (k=40)	11.83(−4.5)	13.08(−13.63)	0.92	0.35	1.33(+0.24)	0.82	4.09(±0.02)
Mirostat (τ=3)	8.14(−8.19)	**23.53**(−3.18)	0.93	0.34	1.30(+0.21)	0.83	4.12(±0.02)
Typical (τ=0.2)	**14.25**(−2.08)	23.51(−3.20)	0.78	0.30	1.27(+0.18)	0.84	**4.15**(±0.02)
Typical (τ=0.95)	11.59(−4.74)	11.77(−14.94)	**0.96**	0.31	1.21(+0.12)	0.84	4.13(±0.02)

- REP (繰り返し) がTop-kなどより少ない
- 人間による評価値は、Typicalが一番高い

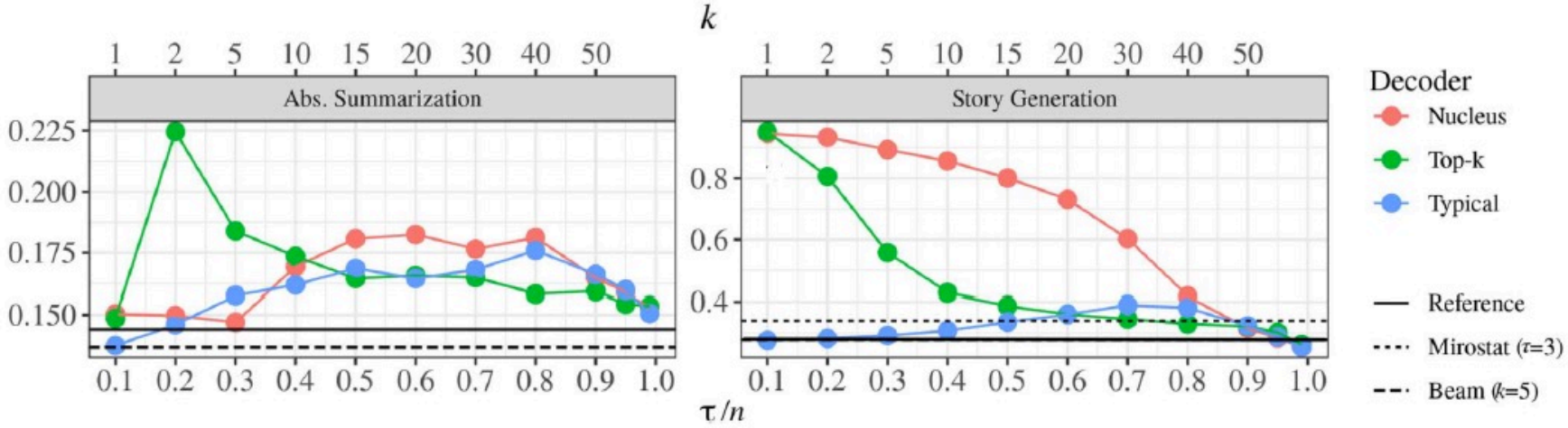
要約タスク

	Abstractive Summarization						
	PPL (g)	PPL (i)	MAUVE (↑)	REP (↓)	Zipf	D (↑)	Human (↑)
Reference	10.29	34.21	–	0.13	0.76	0.97	4.31 (±0.03)
Beam (k=5)	1.39 (−8.90)	**34.21** (−0.00)	0.90	**0.14**	**0.77** (+0.01)	0.97	**4.35** (±0.03)
Temperature (τ=0.5)	**7.10** (−3.19)	55.31 (+21.1)	0.97	0.15	**0.75** (−0.01)	0.97	4.25 (±0.03)
Temperature (τ=1)	6.46 (−3.83)	35.96 (+1.75)	0.95	**0.14**	**0.75** (−0.01)	0.97	4.29 (±0.03)
Nucleus (η=0.9)	2.97 (−7.32)	33.63 (−0.58)	0.90	0.17	0.93 (+0.17)	0.96	4.26 (±0.03)
Nucleus (η=0.95)	3.96 (−6.33)	56.43 (+22.22)	**0.99**	0.15	0.91 (+0.15)	0.97	4.26 (±0.03)
Top-k (k=30)	3.13 (−7.16)	34.79 (+0.58)	0.98	0.16	0.93 (+0.17)	0.97	4.31 (±0.03)
Top-k (k=40)	3.26 (−7.03)	28.38 (−5.83)	0.96	0.16	0.93 (+0.17)	0.97	4.29 (±0.03)
Typical (τ=0.2)	3.80 (−6.49)	62.33 (+28.12)	0.72	**0.14**	0.91 (+0.15)	0.97	4.27 (±0.03)
Typical (τ=0.95)	3.86 (−6.43)	56.67 (+22.46)	0.96	0.15	0.92 (+0.16)	0.97	4.32 (±0.03)

- REPは、変わらず低い
- 温度付きサンプリングの性能が意外と高い

繰り返しの多さ (REP)

- 横軸はハイパーパラメータ、青がTypical sampling、黒が正解の繰り返し率



- 繰り返しは、一般に確率が高い系列
  → 確率の高い語を選ぶサンプリングでは、繰り返しが生じやすい

生成例

Reference	A lawyer for Dr. Anthony Moschetto says the charges against him are baseless. Moschetto, 54, was arrested for selling drugs and weapons, prosecutors say. Authorities allege Moschetto hired accomplices to burn down the practice of former associate.
Beam $k = 5$	Dr. Anthony Moschetto faces criminal solicitation, conspiracy, burglary, arson and weapons charges. ‘‘None of anything in this case has any evidentiary value,’’ his attorney says.
Nucleus $\eta = 0.95$	Dr. Anthony Moschetto, 54, pleaded not guilty to charges Wednesday. Two men – identified as James Chmela and James Kalamaras – were named as accomplices.
Top-k $k = 30$	Dr. Anthony Moschetto is accused of providing police with weapons and prescription drugs. Authorities say he was part of a conspiracy to harm or kill a rival doctor. His attorney calls the allegations against his client ‘‘completely unsubstantiated’’
Typical $\tau = 0.95$	Dr. Anthony Moschetto is charged with crimes including arson, conspiracy, burglary, prescription sale, weapons charges. His attorney says ‘‘none of anything in this case has any evidentiary value’’

- □ はプロンプトから予測できないもの
- τ=0.2でデコーディング (＝確率の高い語を優先して生成)

まとめと感想

- 従来のように確率の高い順に単語を予測集合に加えるのではなく、その文脈での平均予測確率との比で加える、Locally Typical Samplingを提案
- 予測分布のエントロピーが大きい (＝平均予測確率が小さい) ときに、より低い確率の語が予測集合に含まれる
- 局所典型性が成り立つことを実データで確認
  - どんな場合でも一様にそうなのか？
  - 差の分布が左右非対称なのは、単なる数学的なartifactか？

最後に

- 情報理論については、MacKay (2003)は非常に素晴らしい教科書です
- 情報理論を忘れている人が多いので、岩波書店から刊行予定の私の教科書では、基礎を丁寧に説明しています (ご意見募集中) http://chasen.org/~daiti-m/textmodel/