統計的自然言語処理と情報理論

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

情報理論研究会 “若手研究者のための講演会”
2016–12–13(火)
SITA 2016

情報理論と自然言語

- 自然言語と情報理論は, 最初から関係が深い
  - Shannon (1948) "*A mathematical theory of communication*" より:

3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

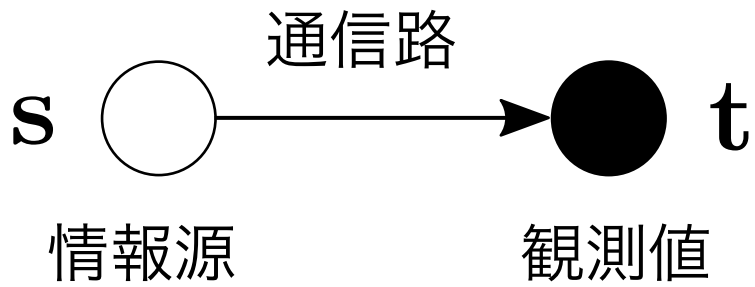
3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

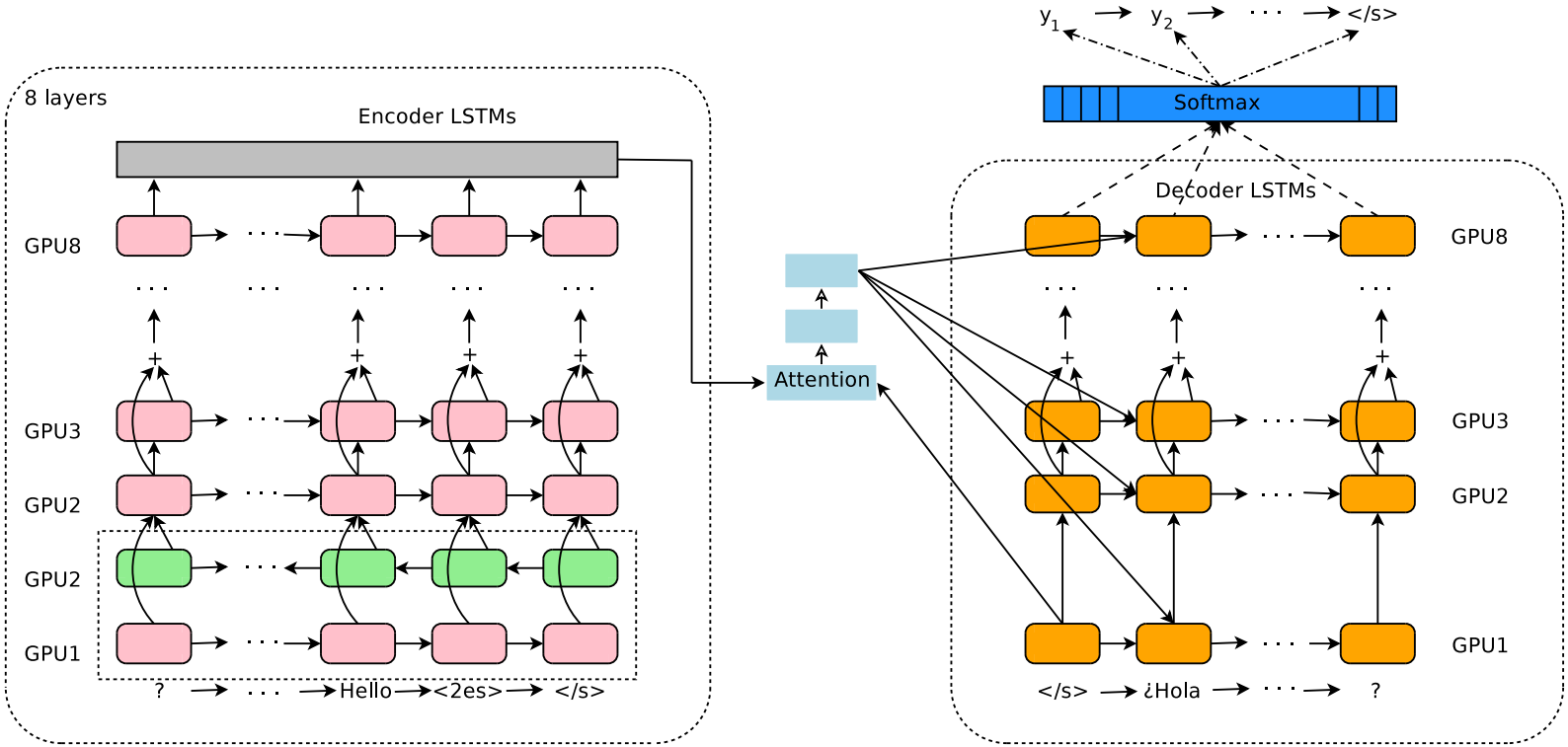
   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

Noisy Channel モデル



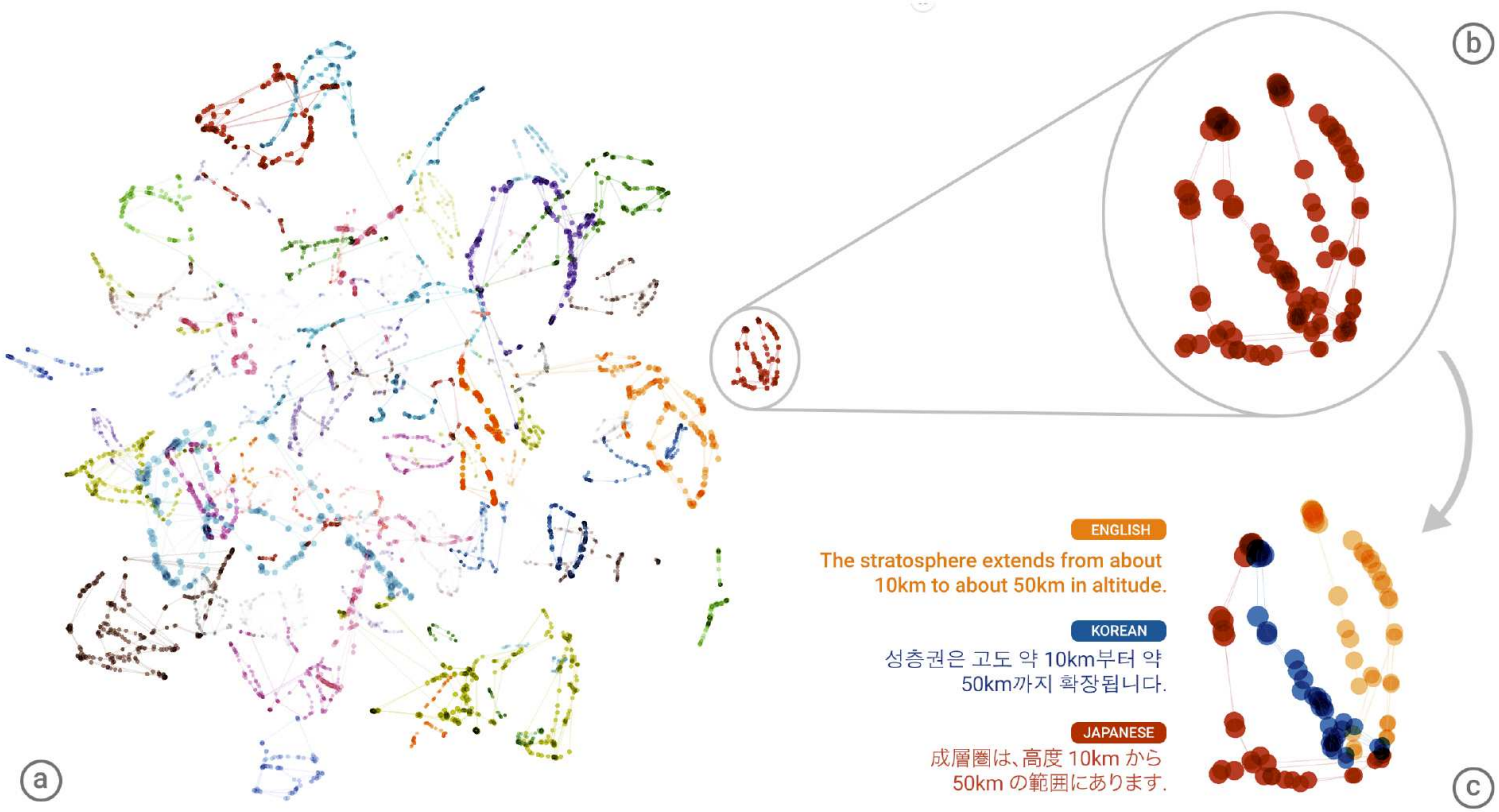
- 自然言語処理のあらゆる場所で使われている
  - 機械翻訳: 言語 A → 言語 B
  - 音声認識: 文 → 音声
  - 表記正規化: 正しい英文 → SNS の崩れた英語 (Eisenstein+ EMNLP2013)

Google Neural 機械翻訳 (2016)



- 2016 年：翻訳精度の大幅な向上 (日本語として読める)
- 原文の「意味」をニューラルネットで数百次元のベクトルに圧縮
  → ニューラルネットで目的言語にデコーディング

文の「意味」の埋め込み



- 楕円の中は「成層圏は, 高度 10km から 50km の範囲にあります」という文の埋め込み
- オレンジ: 英語, 青: 韓国語, 赤: 日本語
- 内部動作のメカニズムは, 未だ不明

情報理論と自然言語処理の違い

- 情報理論…あえて意味を捨象 (実際にはモデル化が必要)
- 自然言語処理…意味を直接扱う (実際には通信理論が必要)

  - ユニバーサル符号である必要はない (「?」→「！」)
  - 言語の知識をどう獲得し, 利用するか?

情報理論と自然言語処理の違い (2)

- データ構造の違い
  - 情報理論は時系列を扱う
    (PPM, HMM, Lempel-Ziv, ···)
  - 自然言語処理は複雑な構造を考える
    (PCFG, 依存構造, トピック, 照応解析, ···)
- アルファベットの違い
  - 情報理論では, アルファベットは 0/1 のことが多い
    (高々 256)
  - 自然言語処理では, 単語種類数は 数万次元以上

情報理論と自然言語処理

講演者 (持橋) の研究での例

- 情報源符号化
  Context-Tree Weighting Method (Willems+ 1985)
  ▽
  無限 Markov モデル (持橋+, NIPS 2007)
- タイプ理論 (Csiszár 1998)
  ▽
  文書や音楽の「典型度」 (中野&持橋+, ISMIR 2016)

CTW法と無限Markovモデル

"*The Infinite Markov Model*", D.Mochihashi, E.Sumita, NIPS 2007. (統計的機械学習)

データ圧縮と統計モデル

- 良い圧縮率を達成することは, データの良い統計モデルを作ることと等価 [韓 94]
- 算術符号化: 文字列 $x_1 \cdots x_{t-1}$ が与えられたとき, 次の文字 x_t の確率

$$p(x_t | x_1 \cdots x_{t-1}) \tag{1}$$

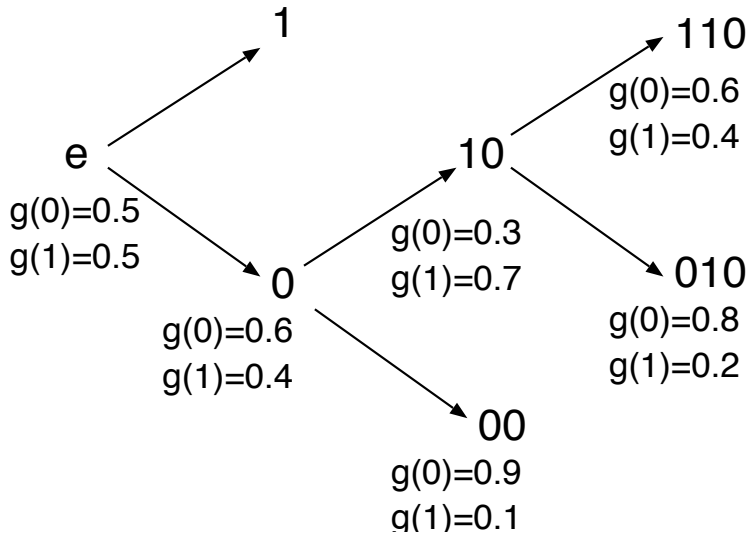
を計算することで符号化

  - PPM-II / PPMd: “escape” でより短い文脈と補間

    ▽

  - Kneser-Ney スムージング (Kneser&Ney 1995) と等価!

Context Tree Weighting method (Willems+ 1995)



- 次の文字の確率を, Suffix Tree 上で再帰的に計算

$$p(x_t|x_1 \cdots x_{t-1}) = \begin{cases} p(x_t|s) & (s \text{ が葉}) \\ \gamma\, p(x_t|s) + (1-\gamma)\, p(x_t|0s)p(x_t|1s) & (\text{otherwise}) \end{cases}$$

- $p(x_t|s)$ は Krichevski-Trofimov (KT) estimator (Be(1/2,1/2))

$$p(x_t|s) = \frac{n(x_t) + 1/2}{n(x_t=0) + n(x_t=1) + 1} \qquad (2)$$

CTW 法の問題点

$$p(x_t|x_1 \cdots x_{t-1}) = \begin{cases} p(x_t|s) & (s \text{ が葉}) \\ \gamma\, p(x_t|s) + (1-\gamma)\, p(x_t|0s)p(x_t|1s) & \\ & (\text{otherwise}) \end{cases}$$

- γ の設定? (γ は均一でよいか?)
- 葉での出力確率? (二値アルファベットでない場合?)
- ヒューリスティックな解決: (岡崎・今井 2000) (定兼+ SITA1999)

n グラムモデル

- n グラムモデル $\cdots$ 言語の予測モデル
  - $p(\text{話す} \mid \text{彼女 が}) = 0.2,\ p(\text{処理} \mid \text{自然 言語}) = 0.7,$ $p(\text{見る} \mid \text{彼女 が}) = 0.1 \ldots$
  - 文の確率を, 予測確率の積に分解 [マルコフ過程] $p(\text{彼女 が 見る 夢}) = p(\text{彼女}) \times p(\text{が} \mid \text{彼女}) \times$ $p(\text{見る} \mid \text{彼女 が}) \times p(\text{夢} \mid \text{が 見る})$
- 各単語は, 前の $(n-1)$ 語の単語のみに依存する

$$p(w_1 \ldots w_T) = \prod_{t=1}^{T} p(w_t | w_{t-1} w_{t-2} \cdots w_1) \tag{3}$$

$$\simeq \prod_{t=1}^{T} p(w_t | \underbrace{w_{t-1} \cdots w_{t-(n-1)}}_{n-1 \text{ 語}}) \tag{4}$$

  - n グラムモデル = 前の $(n-1)$ 語を状態としたマルコフモデル

n グラムモデルの問題

$$p(w_1 \dots w_T) \simeq \prod_{t=1}^{T} p(w_t | \underbrace{w_{t-1} \cdots w_{t-(n-1)}}_{n-1 \text{ 語}}) \tag{5}$$

- n グラムモデル = 単語の総数 V に対して, V^{n-1} 個の状態数
  - 指数的に爆発する
  - $V = 10000$ のとき, $V^2 = 100000000$ (3 グラム), $V^3 = 1000000000000$ (4 グラム), $\cdots$
- 通常, $n = 3 \sim 5$ 程度が限界
  - Google 5 グラムは gzipped 24GB, $V = 13653070$
    - $V^2 = 186406320424900$ (3 グラム)
    - $V^3 =$ (天文学的な数) (4 グラム)
  - しかし, そもそも…

n グラムモデルの問題 (2)

- n グラムモデルは, 単純な $(n-1)$ 次のマルコフ過程
  ＝直前 $(n-1)$ 語を丸覚え
  - 3 グラム, 4 グラム, 5 グラム, ... のデータはノイズだらけ
    - に 英語 が
    - の が # #
    - は 修了 宮本 益
    - が あり 独自 に 法医学
    - は ゼネラル・モーターズ GM や
  - 空間計算量・時間計算量の点でも,
    非常に無駄が大きい
- 言語的に意味がある n グラムは何か?

n グラムモデルの問題 (3)

- 現実の言語データには, 3 グラム, 5 グラムを超えるような長い系列が頻繁に現れる
  - the united states of america
  - 京都 大学 大学院 情報 学 研究科
  - 東京 地検 特捜 部 の 調べ に よる と
  - そんな 事 より 1 よ 、 ちょいと 聞い て くれ よ 。…
- チャンク (句) とみなして一単語にする方法もあるが…
  - 人間の主観的な “正解” に依存
  - どこまでを句とすればよいか [境界は 1/0 か?]
  - 上のような, 慣用句などのフレーズを全て列挙するのは不可能
- バイオインフォマティクス等とも共通する問題
  - DNA, アミノ酸, タンパク質などの系列
    - “正解” が自明ではない

可変長 n グラム言語モデル

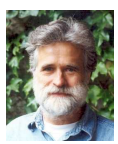
- n グラムの n を文脈に応じて可変長にできないか?
  - “可変長 n グラム言語モデル” … 音声言語分野を中心に提案
  - 踊堂, 中村, 鹿野 (1999), Stolcke (1998), Siu and Ostendorf (2000), Pereira et al. (1995) など

    ⇓
- これまでの “可変長 n グラム言語モデル”= 巨大な n グラムモデルの枝刈りによる方法
  - 指数的に爆発する最大モデルが事前に必要
    - 可変長モデルの意図と矛盾
    - n グラムを減らすことはできても, 増やすことはできない
  - MDL, KL ダイバージェンスなどによる枝刈り
    - 性能があまり悪化しないように減らす
    - 基準はモデルとは別で, 後付け

可変長 n グラム生成モデル

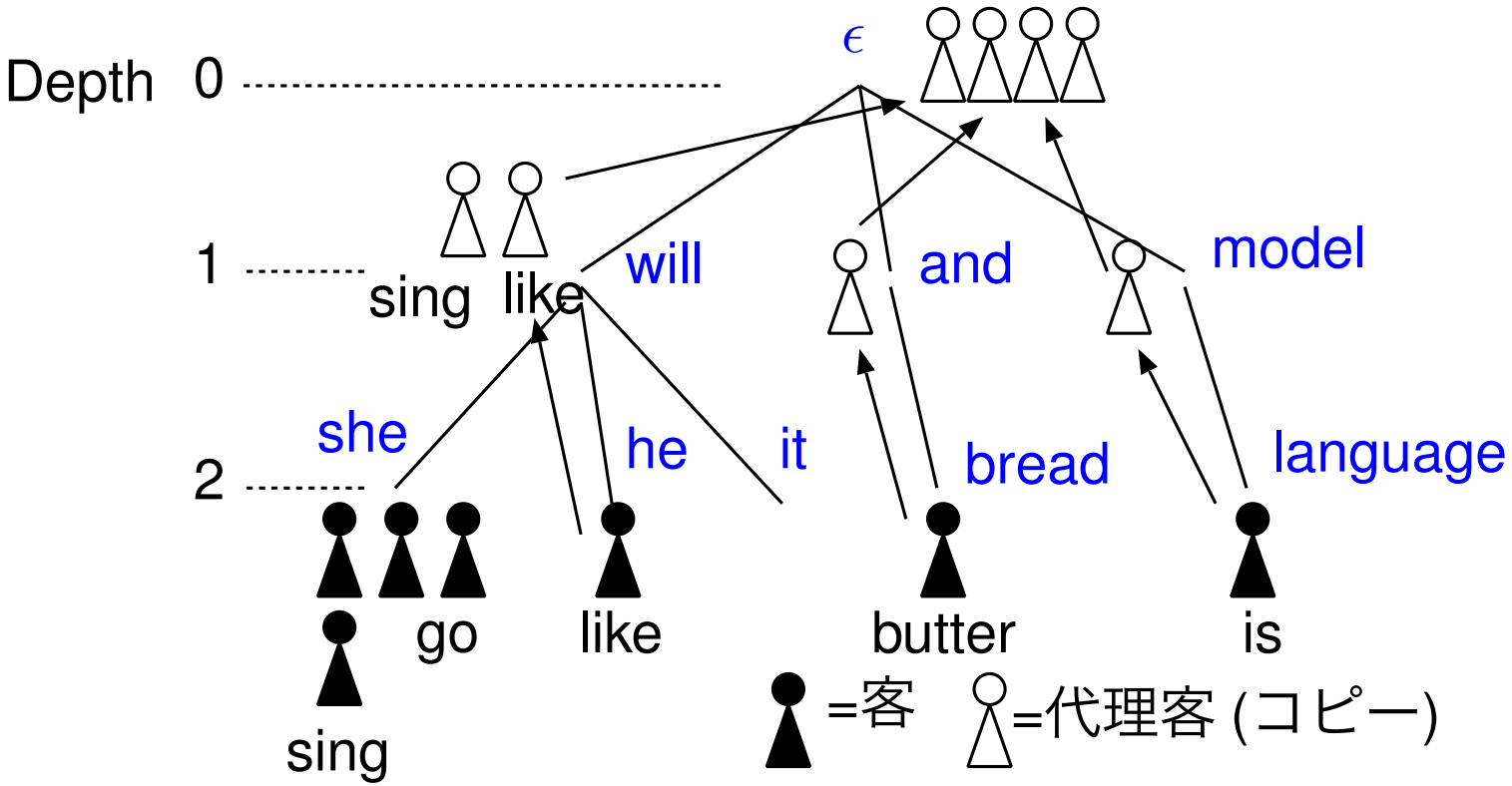
- なぜ, 正しい可変長生成モデルが存在しなかったか?
  $\Downarrow$
  - n グラム分布は, n が大きくなるほどスパース
  - n グラム分布は, $(n-1)$ グラム分布に依存
  - これを階層的に生成する理論的なモデルは存在しなかった
- しかし..

ベイズ nグラム言語モデル

- Hierarchical Pitman-Yor Language Model (HPYLM) (Yee Whye Teh, 2006)
  - 階層ベイズの考えに基づく, nグラム言語モデルの完全なベイズ生成モデル
  - Kneser-Ney スムージングと同等以上の性能 (K-N はその近似)
  - 階層ディリクレ過程 (HDP) の拡張
- Pitman-Yor 過程 (=2-パラメータポアソン=ディリクレ過程 $\mathrm{PD}(\alpha, \theta)$ (Pitman and Yor 1997)) を階層化
  -  Marc Yor (Université Paris VI, France)
  -  Jim Pitman (Dept. of Statistics, Berkeley)

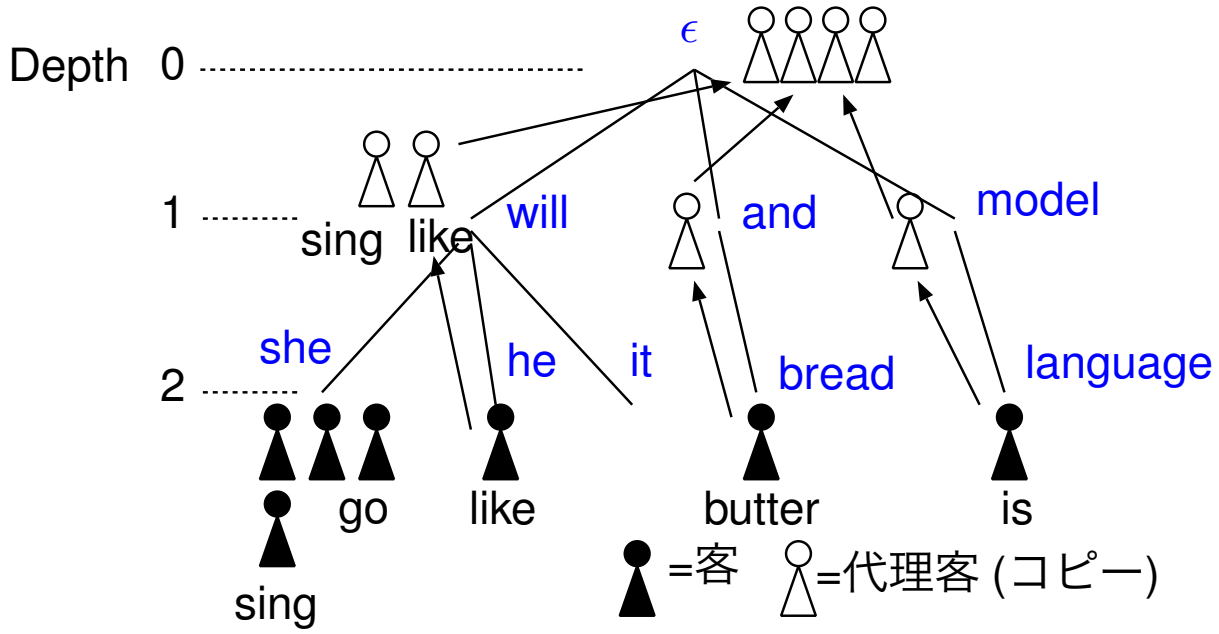
HPYLM (1)

- nグラム分布は, 深さ $(n-1)$ の Suffix Tree で表せる
- 例として, トライグラムを考える



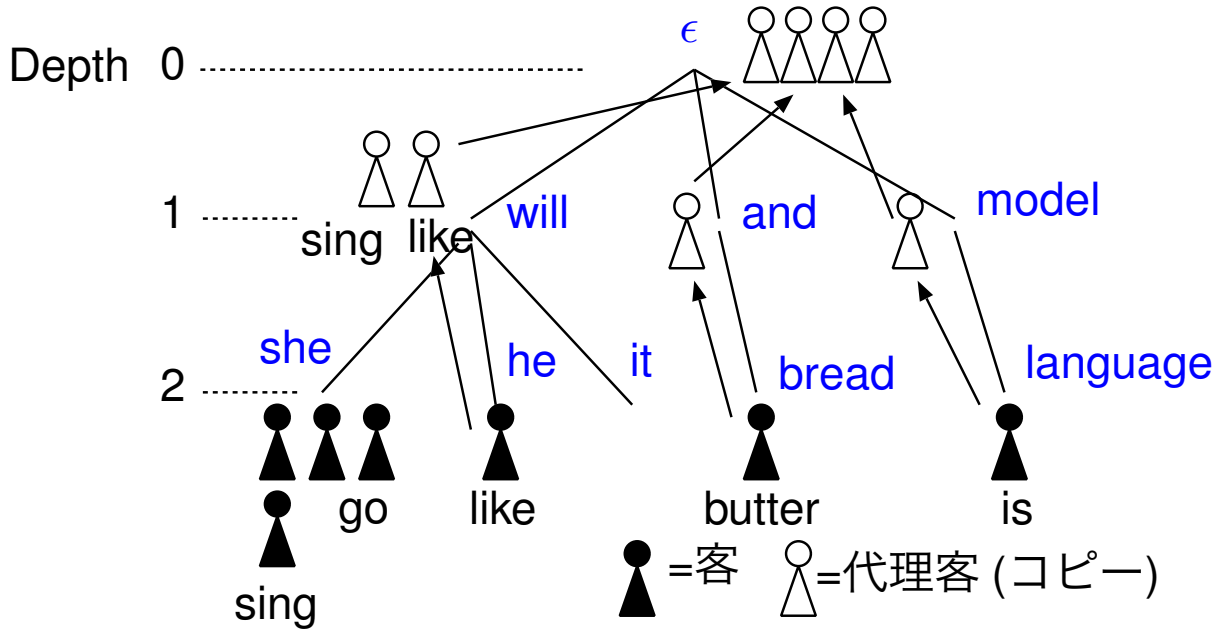
- 'she will'→'sing' を予測…木を $\epsilon \to$ will $\to$ she の順にたどる
- 止まった, 深さ 2 のノード (トライグラム) から, sing の確率を計算

HPYLM (2)



- ノードの持つ客 (単語カウント) の分布から, $p(\text{sing}|\text{she will})$ を計算 $\rightarrow p(\text{like}|\text{she will})$ はどうする?
  - 'like' のカウントがない
- 客のコピー (代理客) を上のノードに確率的に送る
  - 'he will like' から送られた上のノードの客 'like' を使って, バイグラムと補間して確率を計算

HPYLMから可変長モデルへ

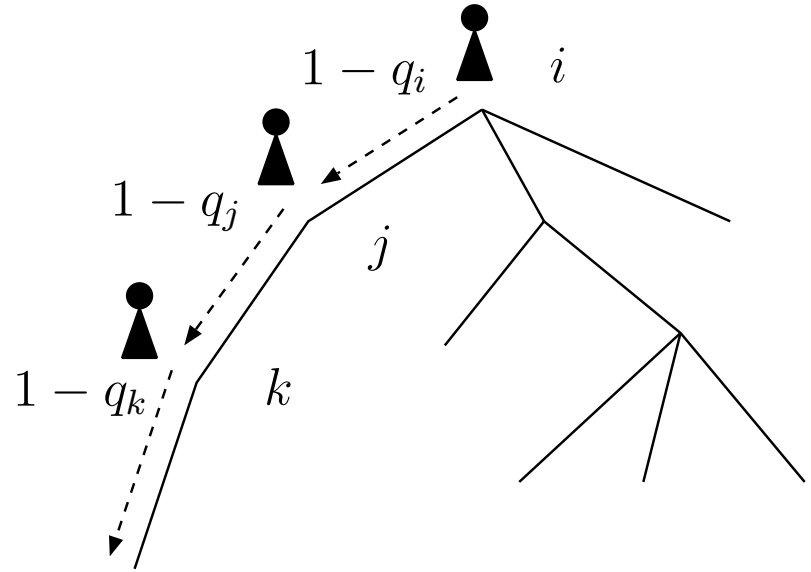


- HPYLM の問題…客 = データのカウントがみな, 深さ 2 のノードに集まるのでいいか?
  - 'will like' は本当は深さ 1 (バイグラム) で十分
  - 'the united states of america'はもっと深いノードが必要

⇓

- 客を違った深さに追加する確率過程.

Variable-order Pitman-Yor Language Model



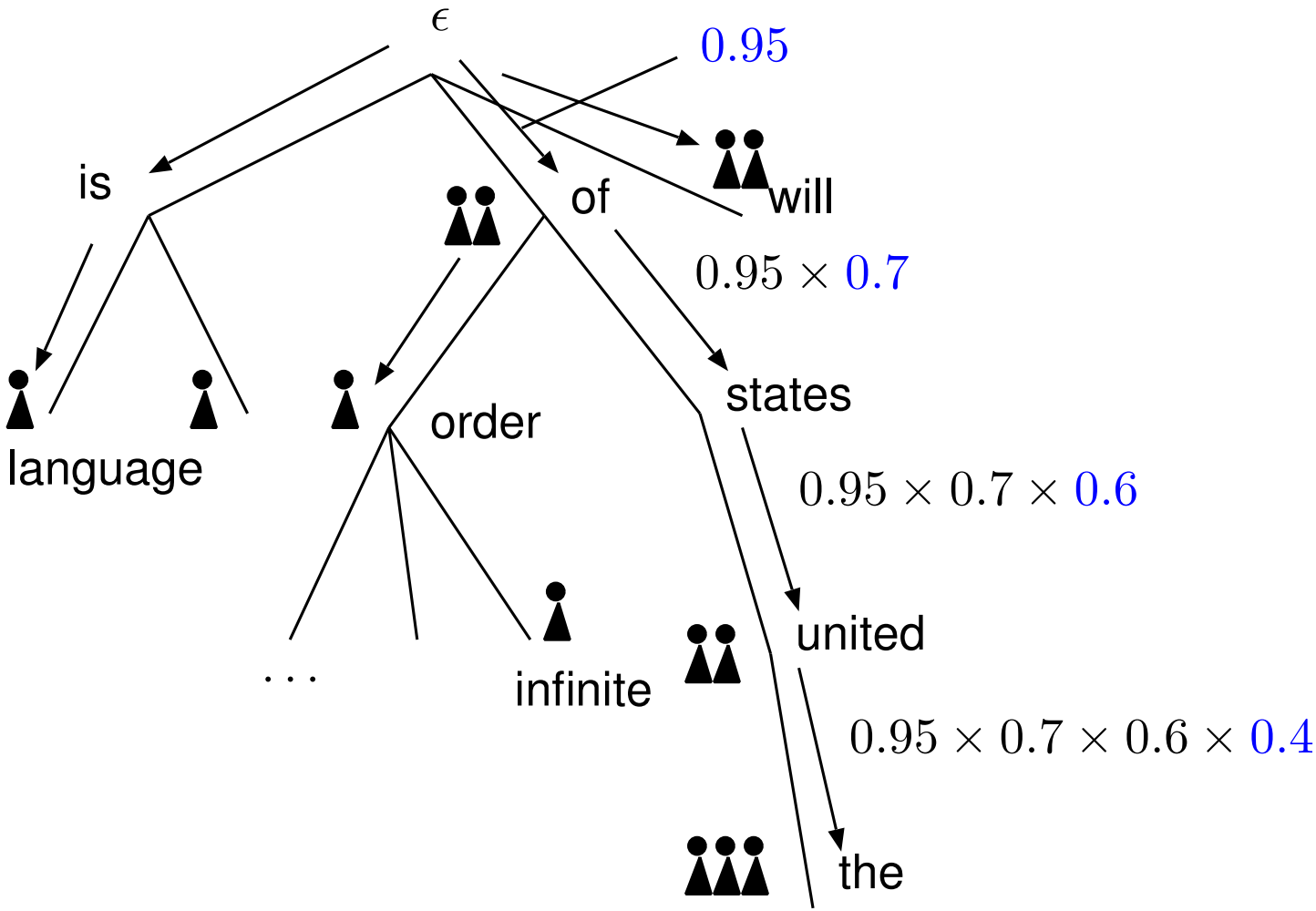
- 客 (カウント) を, 木のルートから確率的にたどって追加
- ノード i に, そこで止まる確率 q_i ($1-q_i$:「通過確率」) がある
  - q_i は, ランダムにベータ事前分布から生成される

$$q_i \sim \mathrm{Be}(\alpha, \beta) \tag{6}$$

  - ゆえに, 深さ n のノードで止まる確率は

$$p(n|h) = q_n \prod_{i=0}^{n-1} (1 - q_i)\,. \tag{7}$$

VPYLM (2)



- 「通過確率」 $1 - q_i$ が大きい … 客が深いノードに到達できる
  - 長い n グラムに対応する
- 「通過確率」が小さい … 'will' など, 浅いノードで充分な文法的

Inference of VPYLM

- もちろん, われわれは自然言語の Suffix Tree がもつ真の q_i の値は知らない
  - どうやって推定したらいい?
- VPYLM の生成モデル: 訓練データ $\mathbf{w} = w_1 w_2 w_3 \cdots w_T$ に,
  隠れた n-gram オーダー $\mathbf{n} = n_1 n_2 n_3 \cdots n_T$ が存在

$$p(\mathbf{w}) = \sum_{\mathbf{n}} \sum_{\mathbf{s}} p(\mathbf{w}, \mathbf{n}, \mathbf{s}) \tag{8}$$

  $\mathbf{s}$: 代理客を含む客全体の配置
- Gibbs サンプリングにより, この $\mathbf{n}$ は推定できる

Inference of VPYLM (2)

- Gibbs サンプリング: マルコフ連鎖モンテカルロ法 (MCMC) の一種
  - 充分サンプリングを繰り返すと, 真の分布に収束
- 単語 w_t の生成された n-gram オーダー n_t を,

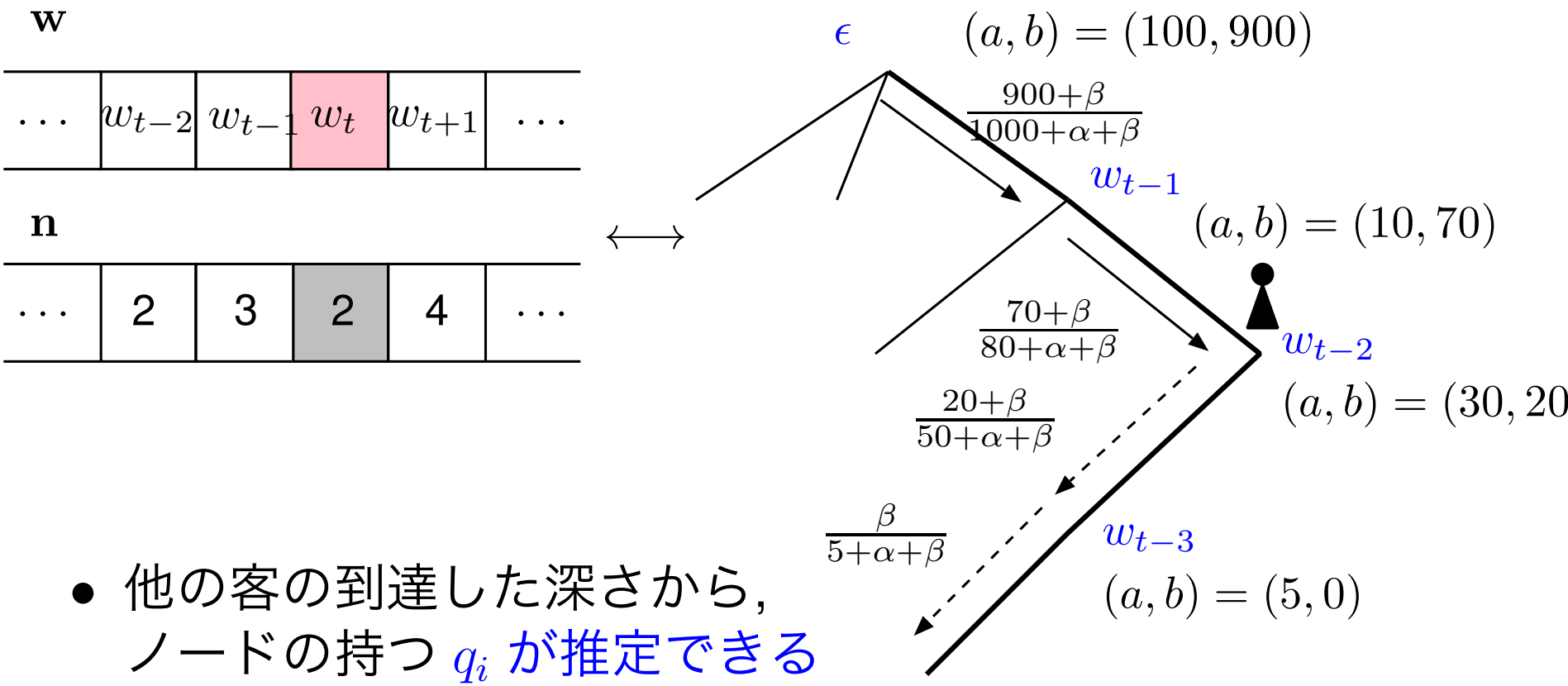
$$n_t \sim p(n_t|\mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \tag{9}$$

  のようにサンプリング
- ベイズの定理より, $n_t = 0, 1, 2, \cdots, \infty$ について

$$p(n_t|\mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) \propto \underbrace{p(w_t|n_t, \mathbf{w}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{n_t\text{-グラムの予測確率}} \cdot \underbrace{p(n_t|\mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t})}_{\text{深さ}n_t\text{に到達する確率}} \tag{10}$$

  - 2 つの確率のトレードオフ (n_t が深すぎるとペナルティ)
  - 第 1 項: HPYLM の n_t-グラム予測確率; 第 2 項は?

Inference of VPYLM (3)



- 他の客の到達した深さから，ノードの持つ q_i が推定できる
- ノード i で他の客が止まった回数を a_i, 通過した回数を b_i とすると，

$$p(n_t = n | \mathbf{w}_{-t}, \mathbf{n}_{-t}, \mathbf{s}_{-t}) = q_n \prod_{i=0}^{n-1} (1 - q_i)$$

$$= \frac{a_n + \alpha}{a_n + b_n + \alpha + \beta} \prod_{i=0}^{n-1} \frac{b_i + \beta}{a_i + b_i + \alpha + \beta}.$$

∞グラム言語モデルの予測確率

- 我々は使うべきnグラムオーダーnを固定しない
  $\rightarrow n$を潜在変数とみなして, 積分消去

$$p(w|h) = \sum_{n=0}^{\infty} p(w, n|h) \tag{11}$$

$$= \sum_{n=0}^{\infty} p(w|n, h)\, p(n|h). \tag{12}$$

- 書き直すと,

∞グラム言語モデルの予測確率

$$p(w|h, n^+) \equiv q_n \cdot \underbrace{p(w|h, n)}_{\text{深さ } n \text{ での予測}} + (1-q_n) \cdot \underbrace{p(w|h, (n+1)^+)}_{\text{深さ } (n+1)+ \text{ での予測}}$$

$$p(w|h) = p(w|h, 0^+)\,,$$
$$q_n \sim \mathrm{Be}(\alpha, \beta)\,.$$

- 無限接尾辞木上の Stick-breaking 過程により, 補間重みを木の場所によってベイズ推定
- CTW で問題だった木の混合比・葉からの予測確率を完全ベイズ化して解決

$$p(x_t|x_1 \cdots x_{t-1}) = \begin{cases} p(x_t|s) & (s \text{ が葉}) \\ \gamma\, p(x_t|s) + (1-\gamma)\, p(x_t|0s) p(x_t|1s) & \\ & (\text{otherwise}) \end{cases}$$

実験

- 英語: 標準的な, NAB (North American Business News) コーパスの Wall Street Journal セットから 10M 語を訓練データ, 1 万文をテストデータ
  - Chen and Goodman (1996), Goodman (2001) などと同じデータ
  - 総語彙数 $=$26,497 語
- 日本語: 毎日新聞データ 2000 年度から, 10M 語 (52 万文) を訓練データ, 1 万文をテストデータ
  - 総語彙数 $=$32,783 語
- $n_{\text{max}} = 3, 5, 7, 8, \infty$ で実験
  - パープレキシティ自体は, $n=7$ 程度で飽和 (Goodman 2001)

テストセットパープテキシティとノード数

(a) NAB コーパス (英語)

n	SRILM	HPYLM	VPYLM	Nodes(H)	Nodes(V)
3	118.91	113.60	113.74	1,417K	1,344K
5	107.99	101.08	101.69	12,699K	7,466K
7	107.24	N/A	100.68	N/A	10,182K
8	107.21	N/A	100.58	N/A	10,434K
∞	—	—	161.68	—	6,837K

(b) 毎日新聞コーパス (日本語)

n	SRILM	HPYLM	VPYLM	Nodes(H)	Nodes(V)
3	84.74	78.06	78.22	1,341K	1,243K
5	77.88	68.36	69.35	12,140K	6,705K
7	77.51	N/A	68.63	N/A	9,134K
8	77.50	N/A	68.60	N/A	9,490K
∞	—	—	141.81	—	5,396K

VPYLM からの生成

「レンタ・カーは空のグラスを手にとり、蛇腹はすっかり暗くなっていた。それはまるで獲物を咀嚼しているようだった。彼は僕と同じようなものですね」と私は言った。「でもあなたはよく女の子に爪切りを買った。そしてその何かを振り払おうとしたが、今では誰にもできやしないのよ。私は長靴を棚の上を乗り越えるようにした。...

- 村上春樹「世界の終りとハードボイルド・ワンダーランド」からのランダムウォーク生成 (VPYLM, $n = 5$)
- 普通の 5-gram LM では, オーバーフィットのため学習データがそのまま生成されてしまう
- 確率的に適切な長さの文脈を用いることで, 特徴をとらえた言語モデル
  - 確率的フレーズ: 『なるほど 」 と 私 は 言っ』(0.6560), 『やれやれ 」 と』(0.7953), 『、と 私 は』(0.8424), . . .

統計的典型度と標準系列

"*Musical Typicality: How Many Similar Songs Exist?* ", T.Nakano, D.Mochihashi, K.Yoshii, M.Goto, ISMIR 2016. (音楽情報処理)

「ありがち」度を測る

- 音楽, 小説, 映画, 文書, …などは爆発的に増えている
(情報爆発)

  ▽

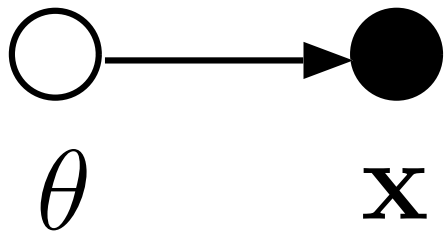
  どれを見るべきか?を知ることが困難
- 多くのデータは「ありがち」(典型的)
  - 「ありがち」でないものが見たい
  - 「ありがち」なものにはどのようなものがあるのか?

  ⇓

  典型性を定量化したい

典型性の定量化

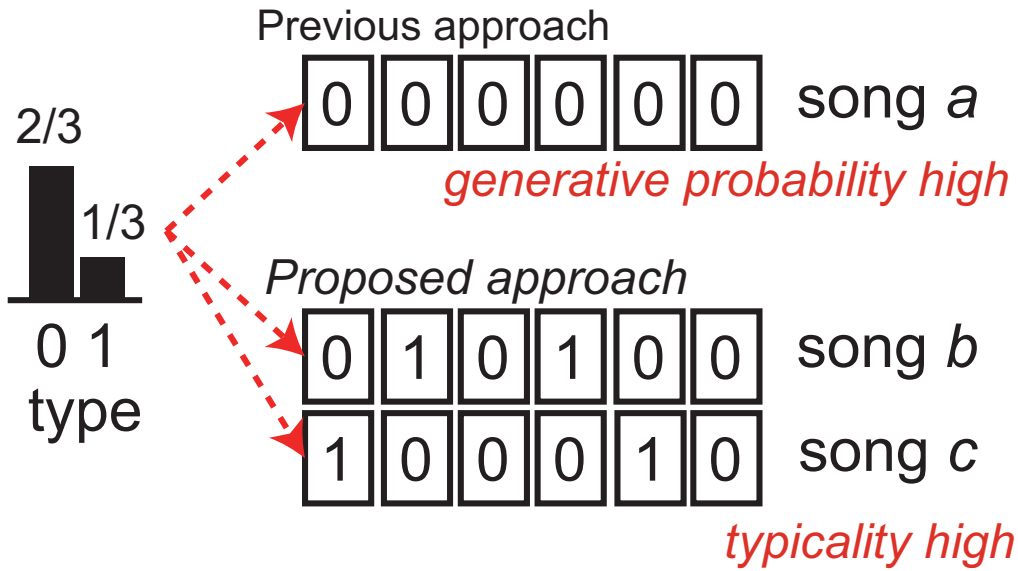
- 確率が高いものが典型的? ([Nakano+ 2015])



θ $\mathbf{x}$

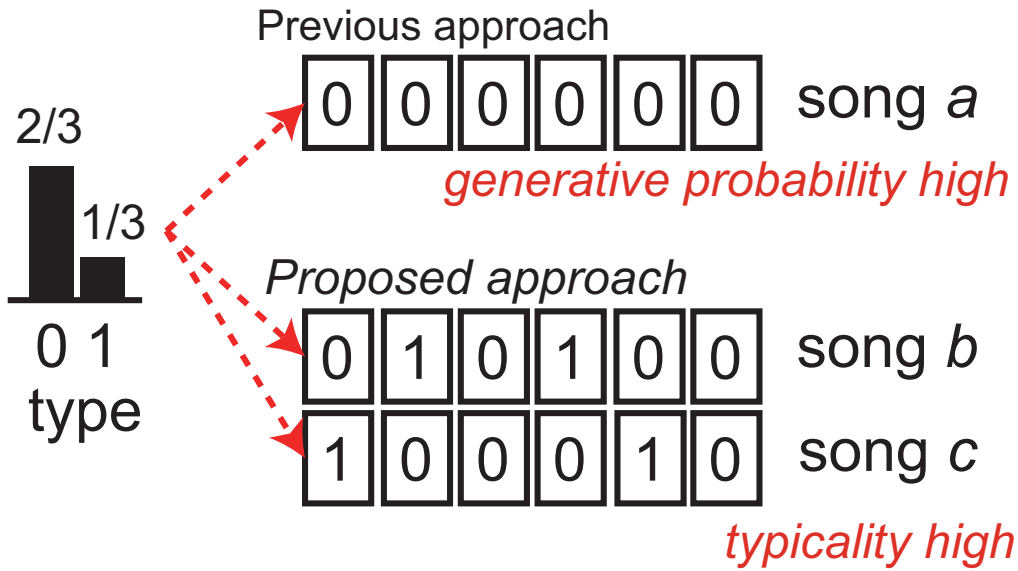
$$\max_{\mathbf{x}} \; p(\mathbf{x}|\theta) \quad ?$$

典型性の定量化 (2)



- そうではない!
- $\{0, 1\}$ を $\left\{\frac{2}{3}, \frac{1}{3}\right\}$ で出す情報源からは, 000000$\cdots$ の確率が最も高くなってしまう
- 言語の例：「のののののののののの」

典型性の定量化 (3)



- 0 と 1 が適度に混ざった 100110001000· · · のような系列が, 典型的な出力のはず

  ⇓

  標準系列! (Typical sequence)

タイプと系列 (Csiszár 1998)

- アルファベット列 $\mathbf{x} = x_1 x_2 \cdots x_n$ ($x_i \in \mathcal{X}$) について, タイプ $P(\mathbf{x})$ とは, 各アルファベットの確率分布 (ここでは経験分布) のこと.

$$P(\mathbf{x}) = \left\{ \frac{1}{n} N(x|\mathbf{x}) \,\middle|\, x \in \mathcal{X} \right\}$$

  - $N(x|\mathbf{x})$: $\mathbf{x}$ の中で x が現れた回数
- 例: $\mathbf{x} = 12243$ のとき,

$$P(\mathbf{x}) = \left\{ \frac{1}{5}, \frac{2}{5}, \frac{1}{5}, \frac{1}{5} \right\}$$

タイプと系列 (2)

- $\mathbf{x} = 000000\cdots$ は確率は高いが, これは 1 通りしかない
- $\mathbf{x} = 101101\cdots$ のような系列は, 多数ある

  $\Downarrow$

  同じタイプを持つ系列の確率の総和が大きい系列が典型的

タイプ Q をもつ情報源が与えられたとき,

1. Q からタイプ P の系列が出現する確率は?
2. タイプ P をもつ系列自体の数は?

タイプと系列 (3)

定理 1

情報源 Q からタイプ P の系列 $\mathbf{x}$ が出力される確率は,

$$Q^n(\mathbf{x}) = \exp\Big[-n\Big(H(P) + D(P||Q)\Big)\Big] \tag{13}$$

Proof.

$$\begin{aligned} Q^n(\mathbf{x}) &= \prod_{i=1}^{n} Q(x_i) = \prod_x Q(x)^{N(x|\mathbf{x})} = \prod_x Q(x)^{nP(x)} \\ &= \prod_x \exp\big[nP(x)\log Q(x)\big] \\ &= \exp\Big[-n\Big(-\sum_x P(x)\log Q(x)\Big)\Big] \\ &= \exp\Big[-n\Big(H(P) + D(P||Q)\Big)\Big]. \quad \square \end{aligned}$$

タイプと系列 (4)

定理 2

タイプ P を持つ系列の集合 $T^n(P)$ の要素数は,

$$\frac{1}{(n+1)^{|\mathcal{X}|-1}} \exp\{nH(P)\} \leq |T^n(P)| \leq \exp\{nH(P)\} \quad (14)$$

Proof: やや複雑なので省略

タイプと系列 (5)

定理 1 と定理 2 から, 情報源 Q の下でタイプ P の系列 $\mathbf{x} = x_1 x_2 \cdots x_n$ の確率の総和は,

$$Q^n(T^n(P)) \doteq \exp(-nD(P||Q)) \tag{15}$$

ただし,

$$a_n \doteq b_n \quad \text{iff} \lim_{n\to\infty} (1/n) \log(a_n/b_n) = 0$$

Proof.

$$\begin{aligned} Q^n(T^n(P)) &= \textstyle\sum_{\mathbf{x}\in T^n(P)} Q^n(\mathbf{x}) \\ &= |T^n(P)| \exp(-n(H(P) + D(P||Q))) \\ &\doteq \exp(nH(P)) \cdot \exp(-n(H(P) + D(P||Q))) \\ &= \exp(-nD(P||Q)) . \quad \square \end{aligned}$$

系列の典型度

$$Q^n(T^n(P)) \doteq \exp(-nD(P||Q))$$

は系列長 n に対して指数的に減少するが(AEP), 我々は n に依存しない性質が知りたいので, パープレキシティと同様に n で割って

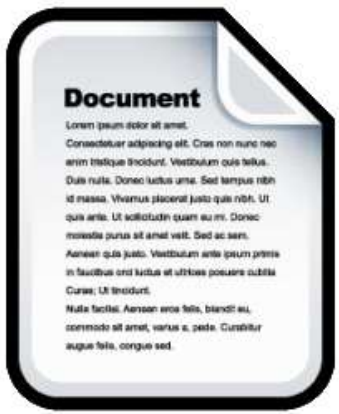
$$\text{Typicality}(P|Q) = \exp(-D(P||Q)) \tag{16}$$

を, 情報源 Q の下でのタイプ P の系列の典型度 (Typicality) と定義する.

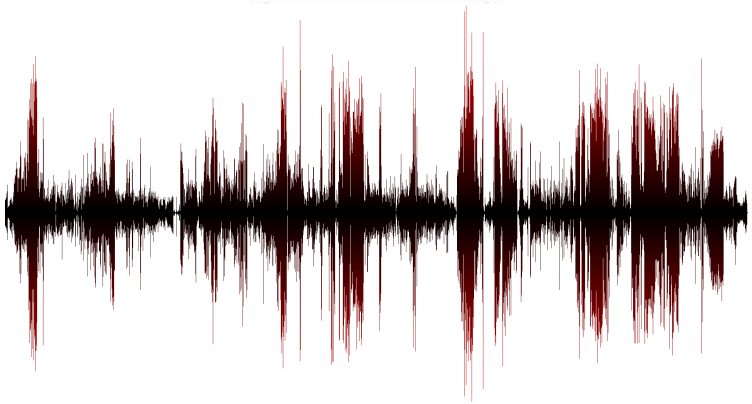
- $\exp(-n\cdots)$ より, $\cdots$ の部分の形に注目している

典型度と言語・音楽

- 実際の言語の単語や音楽の音響データは高次元なので,
  これを潜在的トピックの系列に変換 (LDA; 混合モデル)



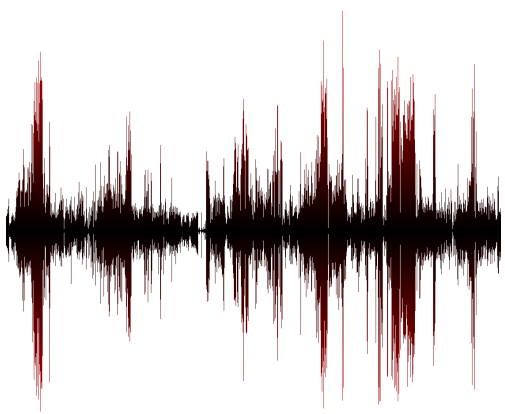
→ 5 3 17 17 2 2 4 · · ·



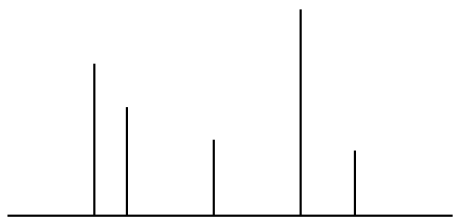
→ 20 5 3 · · · 16 7 2 2 · · ·

典型度と言語・音楽 (2)

- 楽曲を MFCC 系列に変換し, K 平均法でベクトル量子化
  ↓
- 「単語列」だと思って潜在トピックを学習
  ↓
- トピック分布 θ (「タイプ」)



$\rightarrow$ 20 5 3 16 $\cdots$ $\rightarrow$



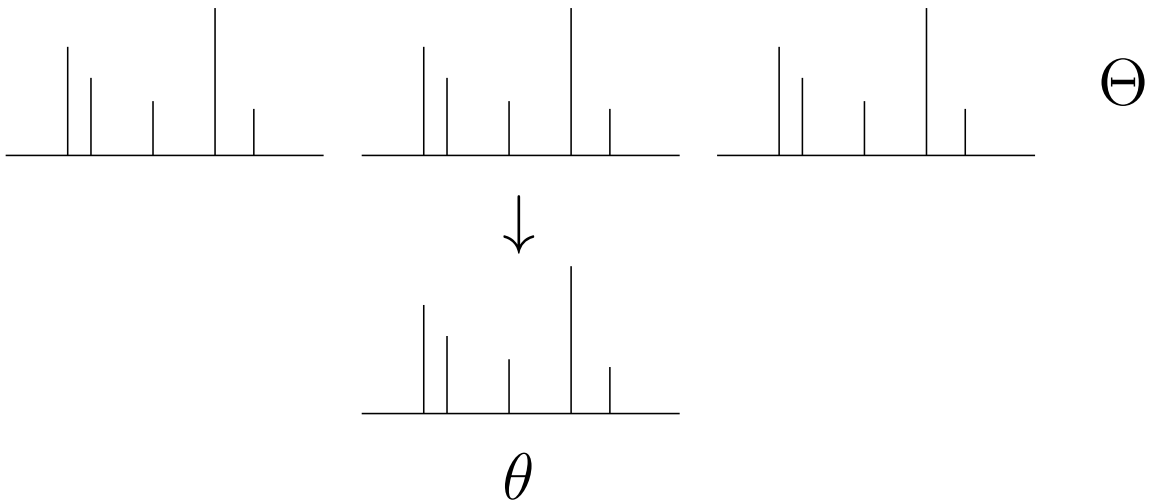
楽曲　　潜在トピック列　　潜在トピック分布 θ

典型度と言語・音楽 (3)

- 楽曲集合とその各曲のトピック分布

$$\Theta = \{\theta_1, \theta_2, \cdots, \theta_M\} \quad (\theta_i : \text{多項分布})$$

が与えられたとき, θ がこの中でどれくらい典型的か?を知りたい

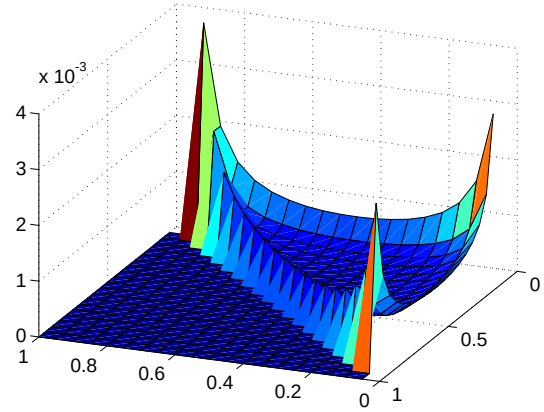


- Θ を生成したディリクレ事前分布 $\mathrm{Dir}(\boldsymbol{\alpha})$ のパラメータ $\boldsymbol{\alpha}$ は, MCMC で推定できる

典型度と言語・音楽 (4)

- 問題: 情報源のトピック事前分布は, 単峰ではない

$$\theta \sim \mathrm{Dir}(\boldsymbol{\alpha})$$



  - $\mathrm{Dir}(\boldsymbol{\alpha})$ の期待値 $\bar{\boldsymbol{\alpha}}$ は, θ を代表しない
  - アルファベット $\mathcal{X}$ は各潜在トピックで, 通常数 100 次元程度・スパース
  - 情報理論では $\mathcal{X} = \{0, 1\}$ なことが多いので, 問題にならなかった

$\Downarrow$

- 情報源のタイプ自体を, 確率的に考える必要

典型度と言語・音楽 (5)

- 情報源のタイプ Q 自体が分布 $\mathrm{Dir}(\boldsymbol{\alpha})$ をもつので, 期待値をとって

$$\mathrm{Typicality}(P|\Theta) = \left\langle \exp(-D(P||\theta)) \right\rangle_{\theta\sim\mathrm{Dir}(\alpha)} \tag{17}$$

$$= \left\langle \exp \sum_{k=1}^{K} p_k \log \frac{\theta_k}{p_k} \right\rangle_{\theta\sim\mathrm{Dir}(\alpha)} \tag{18}$$

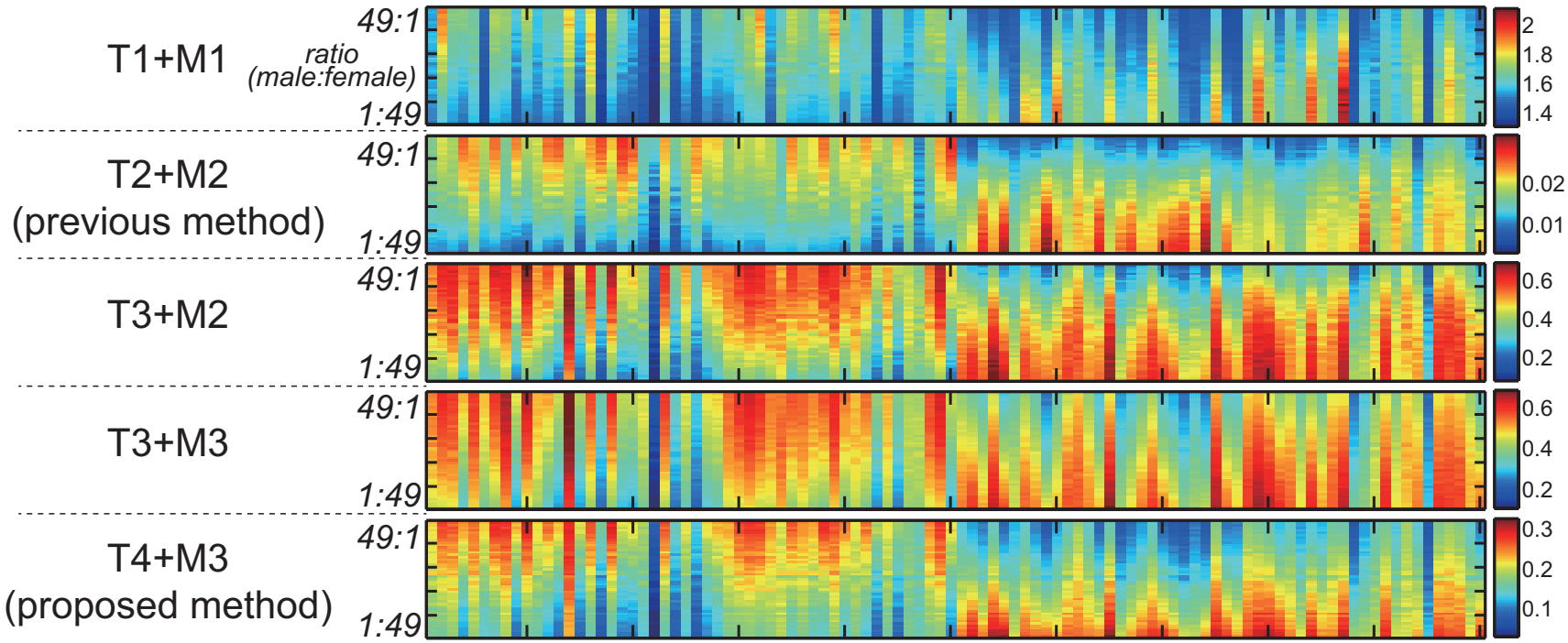
$$= \frac{1}{\exp(\sum_k p_k \log p_k)} \left\langle \exp \sum_k p_k \log \theta_k \right\rangle_{\theta\sim\mathrm{Dir}(\alpha)} \tag{19}$$

$$= \exp(H(P)) \left\langle \prod_{k=1}^{K} \theta_k^{p_k} \right\rangle_{\theta\sim\mathrm{Dir}(\alpha)} = \frac{\exp(H(P))}{\sum_k \alpha_k} \prod_k \frac{\Gamma(\alpha_k + p_k)}{\Gamma(\alpha_k)} \tag{20}$$

実験設定

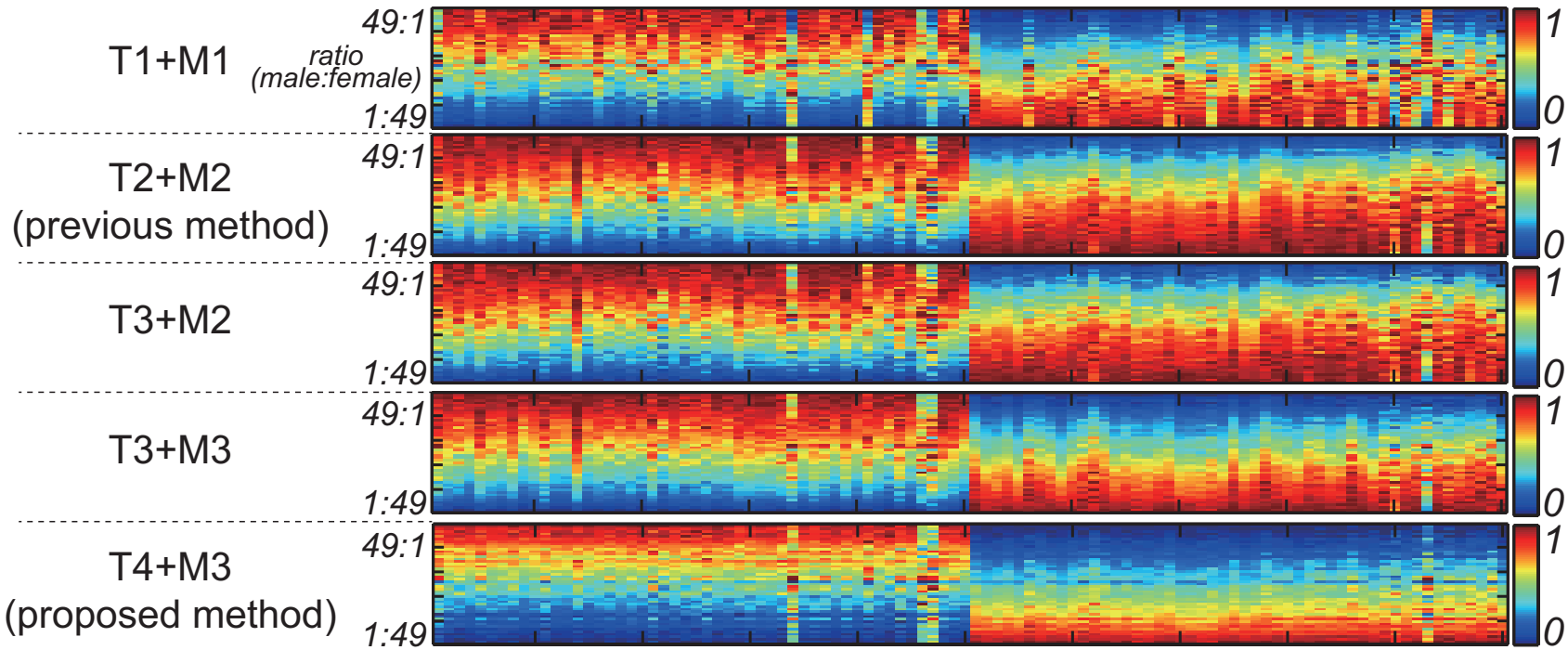
- JPOP MDB: 2000 年-2008 年の期間にオリコンチャートに載った 3,278 曲
- RWC MDB (研究用 100 曲) で音響 GMM を学習して上のデータをベクトル量子化
- LDA 100 トピック ($\mathcal{X} = 100$)
- テスト: 男性ボーカル 50 曲, 女性ボーカル 50 曲
  - 情報源となる 50 曲の男女比を 1:49 $\sim$ 49:1 で変えてテスト
  - 男性曲: 情報源に男性曲が多いほど典型的なはず
  - 女性曲: 情報源に女性曲が多いほど典型的なはず

実験結果 (絶対値)



- 左 50 曲: 男性ボーカル, 右 50 曲: 女性ボーカル
- 各行の縦軸は, 情報源の男女比の割合 (1:49 ～ 49:1)
- 提案法 (最下段) が, 上下がよりはっきり分かれる

実験結果 (相対値)



- 典型度の値を各曲で $[0, 1]$ に正規化
- 最下段の提案法は, 男女の分離が最も明確

統計的自然言語処理と情報理論の今後

- 自然言語処理の社会インフラ化
  - Twitter によるインフル・伝染病・地震などの, テキストによる同期報告:「分散センサ」
  - 多端子情報理論的な設定
- 機械翻訳の実用化: 機械翻訳の「通信路容量」
  - 英語 $\rightarrow$ 日本語でどのくらい情報が失われるか? 中国語 $\rightarrow$ 英語では?
  - 充分に長いメッセージを送れば, 意図している意味が復号できるか?
- 超高次元離散列である言語の中に, 研究のヒントがあるかもしれません

まとめ

- 統計的自然言語処理の特徴と, 情報理論に関係する講演者の研究を紹介した
  - ∞ グラムモデル: CTW 法の拡張+完全ベイズ化 とも見ることができる
  - 文書や音楽の「典型度」: タイプ理論の考え方を高次元離散列に適用
- 自然言語処理は情報理論そのものではないが, 深い関係があり, 情報理論の基本的な考え方を使うことができる
- 今後は, 多端子的な設定や連続量との連係が重要

文献

[1] 韓太舜, 小林欣吾. 情報と符号化の数理 [対象 11]. 岩波講座 応用数学 13. 岩波書店, 1994.

[2] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. In *Proceedings of ICASSP*, volume 1, pages 181–184, 1995.

[3] F.M.J. Willems, Y.M. Shtarkov, and T.J. Tjalkens. The Context-Tree Weighting Method: Basic Properties. *IEEE Trans. on Information Theory*, 41:653–664, 1995.

[4] Daichi Mochihashi and Eiichiro Sumita. The Infinite Markov Model. In *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 1017–1024, 2008.

[5] 持橋大地, 隅田英一郎. Pitman-Yor 過程に基づく可変長 n-gram 言語モデル. 情報処理学会研究報告 *2007-NL-178*, pages 63–70, 2007.

[6] Tomoyasu Nakano, Daichi Mochihashi, Kazuyoshi Yoshii, and Masataka Goto. Musical Typicality: How Many Similar Songs Exist? In *ISMIR 2016*, pages 695–701, 2016.