

A Bayesian Model of Grounded Color Semantics

Brian McMahan, Matthew Stone

統計数理研究所 持橋大地

daichi@ism.ac.jp

最先端NLP勉強会 2015-8-30 (Sun)

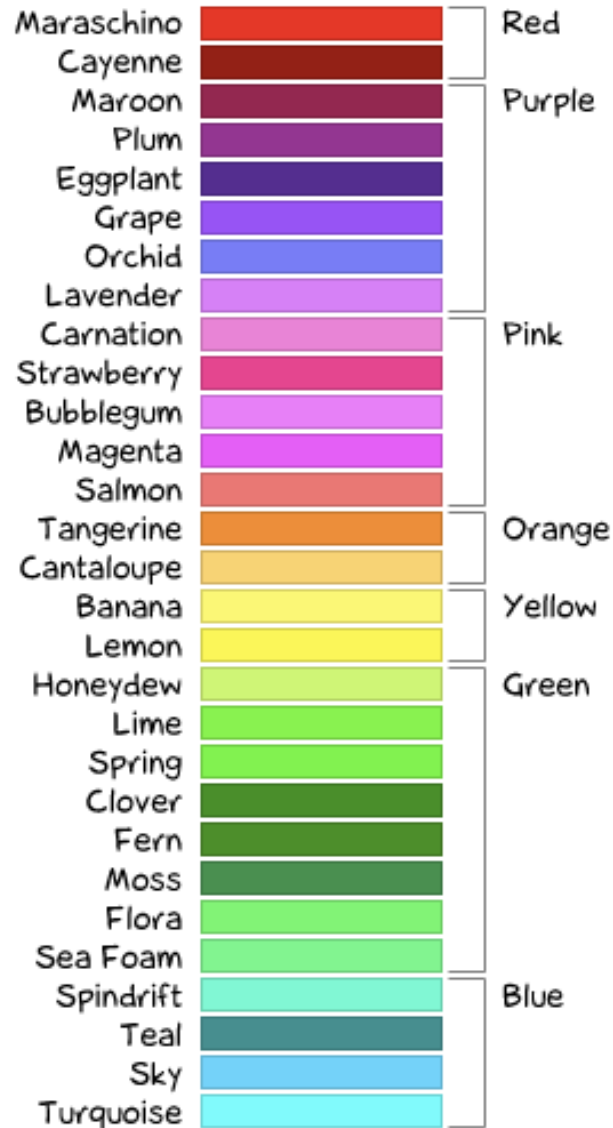
この論文の内容

- 色のRGB値と、その色を表す言葉の対応の学習
 - この論文では色を対象にしているが、同様の話は
 - 量を表す言葉
 - 空間を表す言葉
 - 時間を表す言葉などにも関係が深い
- 研究の動機は面白いが、技術的には微妙
- もっと自然で良いモデル化がありそう

Munroe Color Corpus

- xkcdのエントリ
<http://blog.xkcd.com/2010/05/03/color-survey-results/>
- クラウドソーシングで、与えられた色を表す言葉を書かせる
- データが上のページで公開されている
 - 829個の色
 - 2176417個の言葉

Color names if
you're a girl...



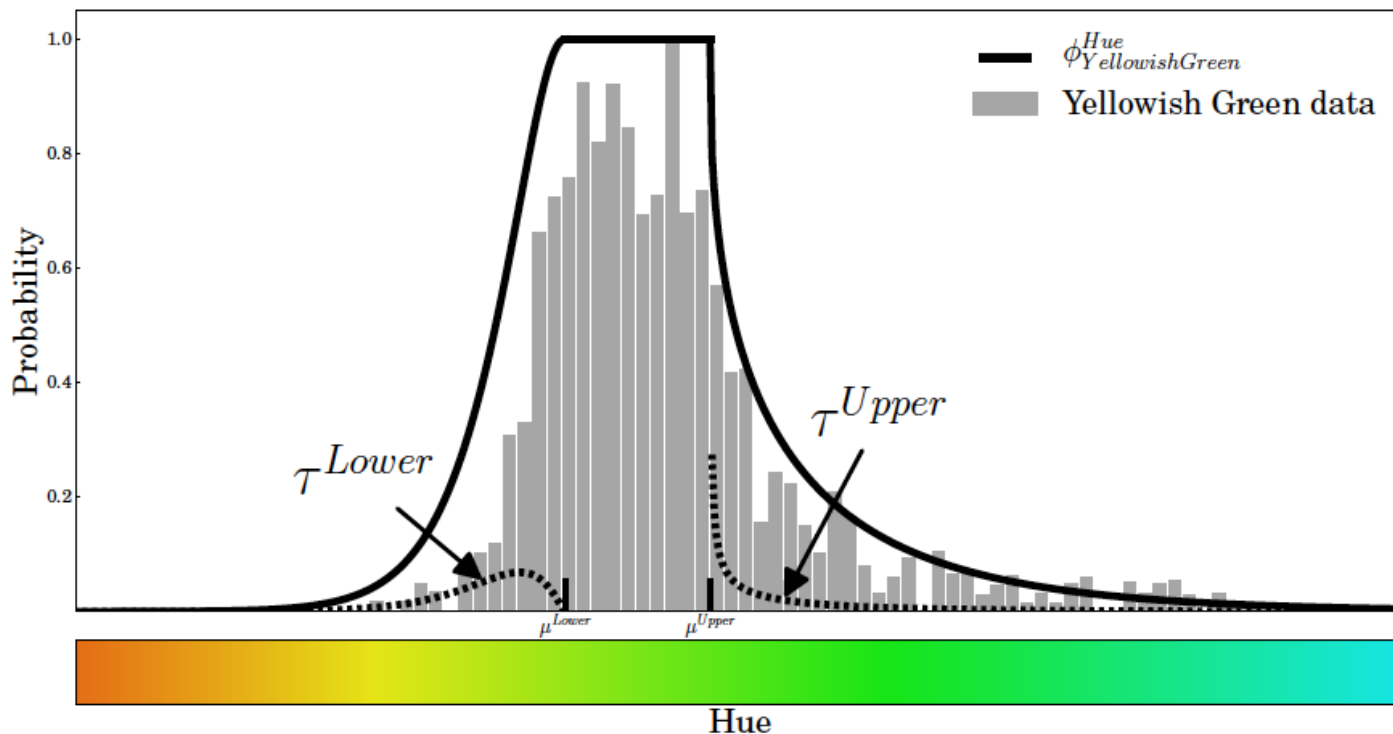
Color names if
you're a guy...

Doghouse Diaries
"We take no as an answer."

色のモデル化

- RGBではなく、HSV (Hue-Saturation-Value)の3次元で表現
 - Hue: 色合い
 - Saturation: 彩度; グレーと比べた鮮やかさ
 - Value: 明度; 色を白-黒でみたときの明るさ

色のモデル化 (2)



- 各次元で、色を上限を下限の間として表現
- 上限=最低値+ガンマ分布
- 下限=最高値-ガンマ分布

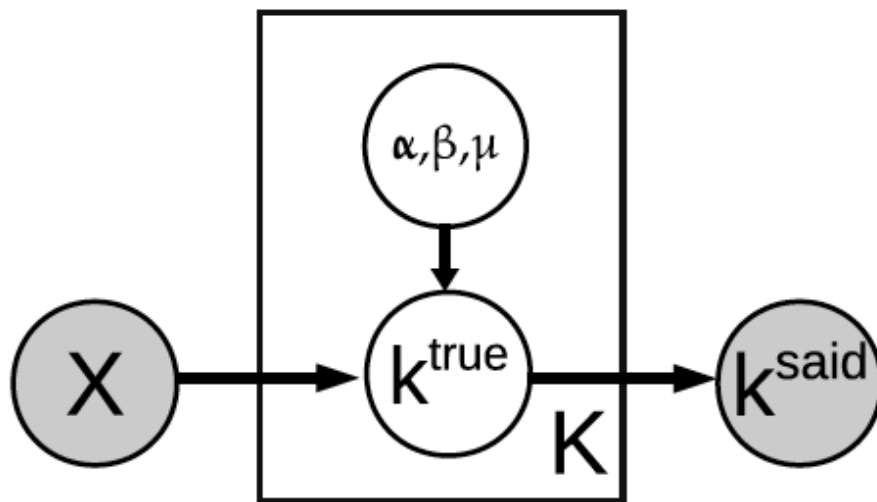
色のモデル化 (3)

- 色の確率=HSVの各値が範囲に入っている確率

$$\begin{aligned} & P(\tau_k^{Lower, H} < x^H < \tau_k^{Upper, H}) \times \\ & P(\tau_k^{Lower, S} < x^S < \tau_k^{Upper, S}) \times \\ & P(\tau_k^{Lower, V} < x^V < \tau_k^{Upper, V}) \\ & = \prod_d P(\tau_k^{L,d} < x_i^d < \tau_k^{U,d}) \end{aligned}$$

- なぜ、値の分布を直接モデル化しないのか謎
 - 言語学者なので、分類思考に毒されている
 - ベータ分布(の混合分布)を使えばOK

色を表す言葉の推定



$$P(k^{\text{said}}, k^{\text{true}} | x) = P(k^{\text{said}} | k^{\text{true}}) P(k^{\text{true}} | x)$$

- “真の言葉”というartifactを考えている
 - こんなものはなくてok
 - 条件つき確率 $p(\text{said}|\text{true})$ は正しくモデル化されない

より良いモデル

- 単純なベイズの定理

$$p(k^{\text{said}} | x) \propto p(x | k^{\text{said}}) p(k^{\text{said}})$$

- 言葉kから色xが出る確率が高くても(=kによる記述が正確でも)、kの頻度が高くないと選ばれない
 - “Chiffon” “Maroon” 等は男性は喋らない
 - 「その人に合わせた色言葉の提案」の必要

実験

- データを70-5-25%に分けて、色を予測できるかどうか試す
- 比較したモデル
 - LUX: 閾値に基づく提案法
 - HM: ヒストグラムモデル. HSV空間の各セルで、そこから出やすい言葉をヒストグラムに
 - スムージングが必須. 3段階の粗視化
 - GM: Gaussianモデル. 各色を、HSV空間上の中心と分散をもつガウス分布に対応づける

実験結果

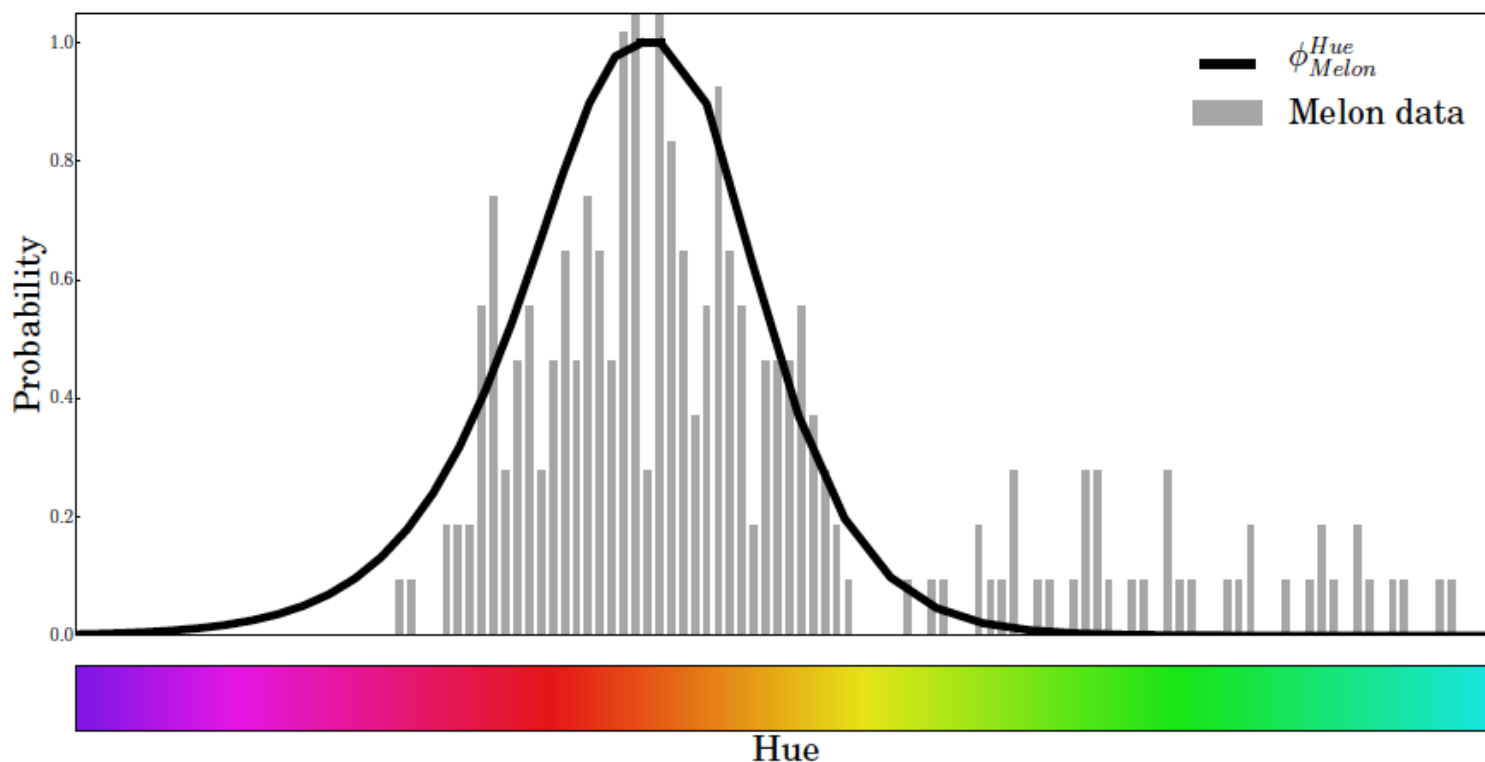
	TOP^1	TOP^5	TOP^{10}
LUX	39.55%	69.80%	80.46%
HM	39.40%	71.89%	82.53%
GM	39.05%	69.25%	79.99%

- ヒストグラムやガウス分布とほとんど変わらない結果. 尤度↓

	$-LL$	$-LLV$	AIC	Perp
LUX	1.13×10^7	2.05×10^6	4.13×10^6	13.61
HM	1.13×10^7	2.09×10^6	4.82×10^6	14.41
GM	1.34×10^7	2.08×10^6	4.17×10^6	14.14

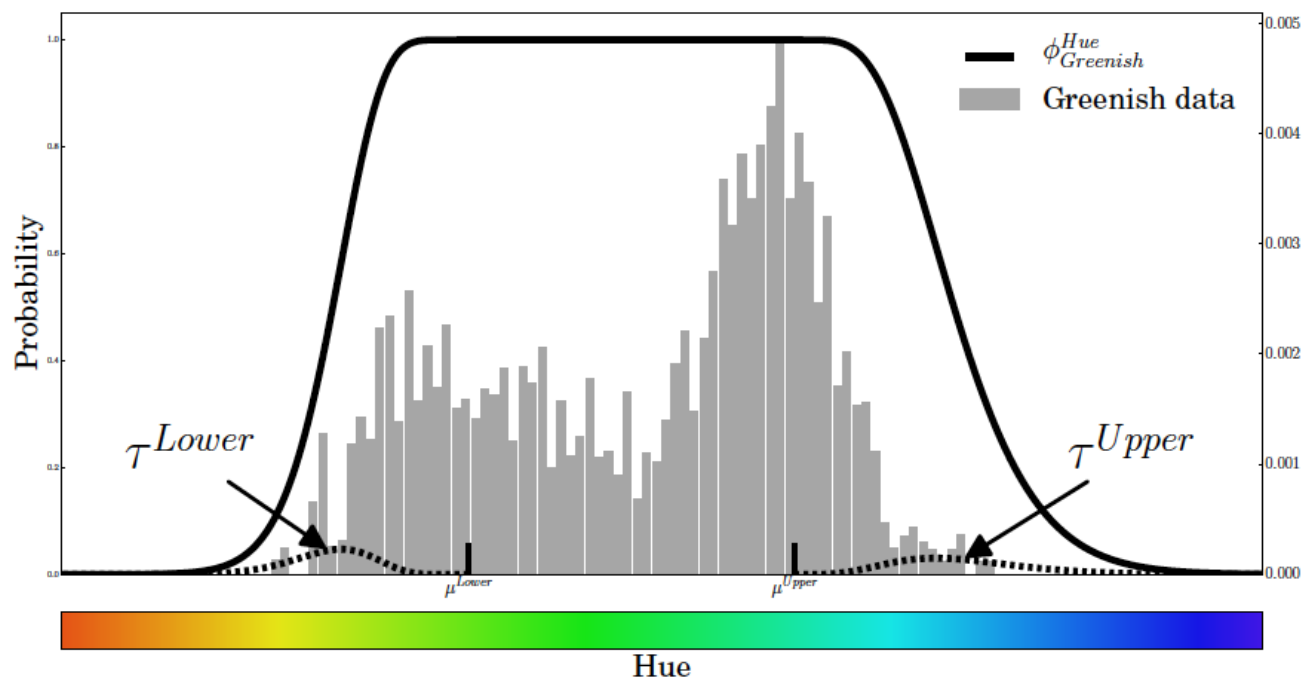
Insufficiency

- “melon”を表す色の分布. 黒線=LUXモデル
 - 長い裾が表現できない (閾値ベースだから)



Insufficiency (2)

- “greenish”を表す色の分布 (黒線=LUX)
 - Multimodalな分布はまったく表現できない



- Dirichlet process mixture of Beta?

結論

- 言語学的に議論を固めてあるが、技術的には疑問
 - より自然で簡単なモデルと、同程度の精度
- データセットの紹介と、色のモデル化の意義を伝えたことには貢献
- これからの課題：
 - 個人に合わせた色の言葉の提案
 - 環境による色の変化にどう対応するか？
 - 朝の「黄色」と夕方の「黄色」は異なる
 - 色空間の全体的なシフト
 - 言葉が表す色の多峰性の同定 など