

Emergence of Syntax Needs Minimal Supervision

Raphaël Bailly, Kata Gábor, ACL 2020 (事前投票4票)

読む人：吉川将司 (東北大), 2020/09/26, 第12回最先端NLP勉強会

注

- ・ 数式が煩雑になるのを避けるため言語の文は常に3語と仮定して書きます (実験も 3 語言語)
- ・ 記法を論文から勝手に変えたりしてます

(概要)

背景と目的：NN言語モデルは統語論を捉えるか

- 系列型の言語モデルも挙動を見ると統語的規則性を捉えてるよ派
人称の文法性判断 [Linzen+, 2016]
- 明示的な教師信号や仮説空間の制限が必要だよ派
再帰型、パーザ×言語モデル (RNNG) [Dyer+, 2016]
- こちらの問題点：

容認性が異なる最小ペア
が用意しにくい

言語モデルによる確率の
僅差は本当に有意か

等々方法に限界があり
- 統語以外のcueを活用してる可能性もあり [Gulordava+, 2018]
e.g. 項の典型性：*dogs that love their friend bark*
- **本研究**：系列型モデルから**直接的に統語情報**（≡品詞）を取り出す手法を提案
Emergence of Syntax
- 「（教師なし付与した）タグ列が統語的」を情報理論の言葉で表現
- 「統語的」なタグセットの空間から最適なものを探す
Minimal Supervision

2大潮流？

ただし実験はまだ予備的

アイデア：文脈的/統語的分割の形式的定義

- 文の形を決定する要因を大別して以下の2つと考える
 - **統語論**：統語論の自律性 [Chomsky 57]
 - この観点における文のwell-formednessは他の要因から独立
 - 古典的な例：*Colorless green ideas sleep furiously*
 - それ以外の**意味、語用論的**（まとめて**文脈的**）な要因
 - 分布仮説 [Harris 54, Firth 57] やトピックモデル等により形式化してきた
 - 同じ文脈、文、文章などの**近接性**によるモデリング
- ⇒ 後者をまず形式的に定義して、統語的要因を「そうでないもの」として定義できないか？
これにより「統語的」を（情報理論の）最小限の言葉で記述できる？

確率的言語 L と語彙 V は固定して考えます

語彙 V の分割

便宜的に要素をカテゴリと呼びます

・ (確率的) 分割 $P = (C, \{\pi_v\}_{v \in V})$

語 v に条件付けられた C 上の確率分布

・ 単語に品詞を付与するイメージ

・ 端的に、探索して統語的性質を捉えた分割を見つけることが目的

・ 分割の性質の記述は L の分割による像、カテゴリ言語 $\pi(L)$ を介して行う

$$p_{\pi(L)}(c_1 c_2 c_3) = \sum_{v_1, v_2, v_3} \pi_{v_1}(c_1) \pi_{v_2}(c_2) \pi_{v_3}(c_3) p_L(v_1 v_2 v_3)$$

$p=1/4$

$p=1/4$ s_1 : cats eat rats

$p=1/4$ s_2 : rats fear cats

s_3 : mathematicians prove theorems

$p=1/4$ s_4 : doctors heal wounds

N V N

品詞によるone-hotな分割 $C=\{N, V\}$

cats eat rats

rats fear cats

mathematicians prove theorems

doctors heal wounds

$p=1/2$
 $p=1/4$
 $p=1/4$

文のトピックによるone-hotな分割 $C=\{\text{🐶}, \text{🤔}, \text{🚑}\}$

文脈的分割

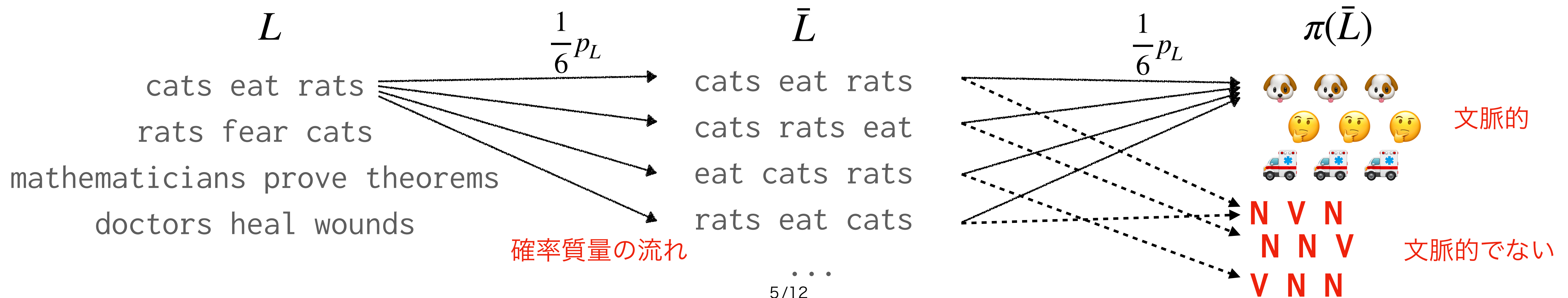
- \bar{L} : $s \in L$ の語順を並べ替えたものをすべて集めた確率的言語

- $p_{\bar{L}}$ を p_L から導出: $p_{\bar{L}}(v_1 v_2 v_3) = \frac{1}{3!} \sum_{(i_1, i_2, i_3) \in \sigma(3)} p_L(v_{i_1} v_{i_2} v_{i_3})$ おおざっぱには6等分して分配

- 定義①: 分割 P が L に対して文脈的 $\stackrel{\text{def}}{\Leftrightarrow} \pi(L) = \pi(\bar{L})$

- 気持ち: 語順情報を壊す前後でカテゴリ列が等確率 $\Rightarrow P$ は語順情報を持たない

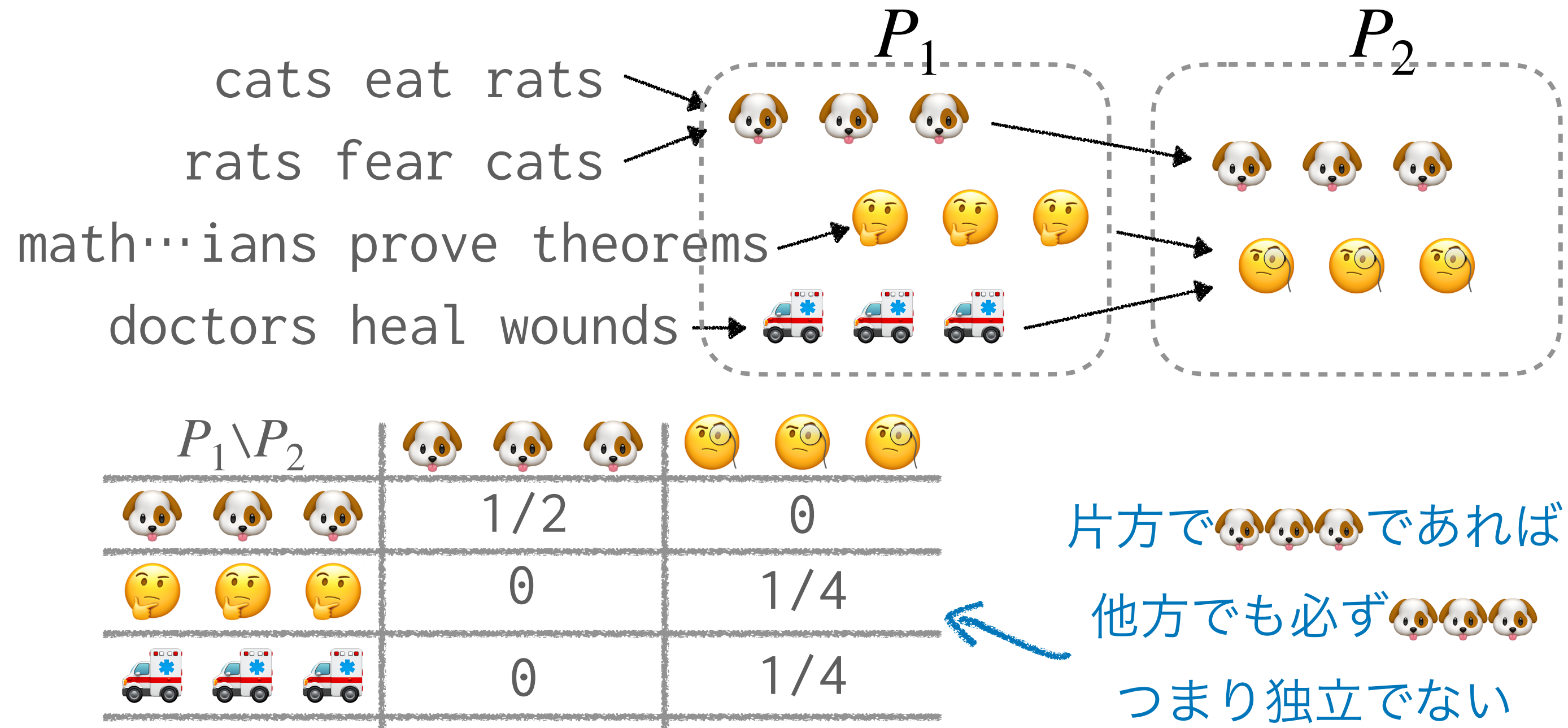
特に、トピックによる分割は常に文脈的 (近接性)。このような分割を包含する概念




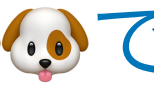




同時分割と分割の独立性










・ 2つの分割の同時分割 $P \cdot P' = (C \times C', \pi \cdot \pi')$, ここで $(\pi \cdot \pi')_v(c, c') \stackrel{\text{def}}{=} \pi_v(c)\pi'_v(c')$

- ・ 定義②：分割 P と P' が**独立**とは、対応するカテゴリ列の言語で確率的に独立
 具体的に P, P' , また $P \cdot P'$ の像言語にて $p_{\pi \cdot \pi'(L)}((c_1, c'_1)(c_2, c'_2)(c_3, c'_3)) = p_{\pi(L)}(c_1 c_2 c_3) p_{\pi'(L)}(c'_1 c'_2 c'_3)$
- ・ 片方のカテゴリ列を知っても他方の列についてはわからない



カテゴリ言語の同時確率分布表

片方で    であれば
 他方でも必ず   
 つまり独立でない

	N	V	N
  	1/2		
  	1/4		
  	1/4		










N V Nを知ってもトピックについては何もわからない！
 つまり独立

統語的分割とはつまり

・ 定義③：分割 P が**統語的**とは、 P が任意の文脈的分割 P' に独立なことをいう

・ 統語的要因を「文脈的でないもの」として形式的に定義できた😊

🤔 1. この定義で直感的な統語論を捉えられるのか？

	N	V	N
  	1/2		
  	1/4		
  	1/4		

⇒ 最後の実験的に示す

🤔 2. 「任意の P' に独立」というのは結構厳しい条件

⇒ 次に情報理論の言葉に翻訳し、その上でrelaxationを行う

情報理論の言葉に翻訳, Relaxation

L における頻度

- ・ 構造なし言語 \bar{L} : L と同じ文集合、確率は単語の頻度 $p_{\bar{L}}(v_1 v_2 v_3) = p_L(v_1) p_L(v_2) p_L(v_3)$

単語同士の共起情報が消え、統語的/文脈的性質がわからない

- ・ 分割 P の情報量 : $I_L(P) = H(\pi(\bar{L})) - H(\pi(L))$
構造なし言語に対するエントロピーの減少量

$$- \sum_{s \in \pi(L)} p_{\pi(L)}(s) \log p_{\pi(L)}(s)$$

- ・ P が文脈的 $\stackrel{\text{def}}{\Leftrightarrow} I_L(P) = I_{\bar{L}}(P)$

$$\because \pi(L) = \pi(\bar{L}) \Leftrightarrow I_L(P) = I_{\bar{L}}(P)$$

緩和
→

$$P \text{ が } \gamma \text{ 文脈的 } \stackrel{\text{def}}{\Leftrightarrow} \min_P I_L(P)(1 - \gamma) - I_{\bar{L}}(P) \text{ の解}$$

- ・ P が統語的 $\stackrel{\text{def}}{\Leftrightarrow}$ 任意の文脈的 P' に対して相互情報量 $I_L(P; P')$ が 0

$$\because P \text{ と } P' \text{ が独立 } \Leftrightarrow I_L(P; P') = 0$$

$$H(\pi(L)) + H(\pi'(L)) - H((\pi \cdot \pi')(L))$$

緩和
→

$$P \text{ が } \mu, \gamma \text{ 統語的 } \stackrel{\text{def}}{\Leftrightarrow} \min_P \max_{P^*} I_L(P; P^*) - \mu I_L(P) \text{ の解 } (P^* \text{ は } \gamma \text{ 文脈的分割})$$

疑問 : L に対して P^* は
ユニークでない?

8/12

これがないと 1 カテゴリーだけの
trivial な解に行ってしまう?

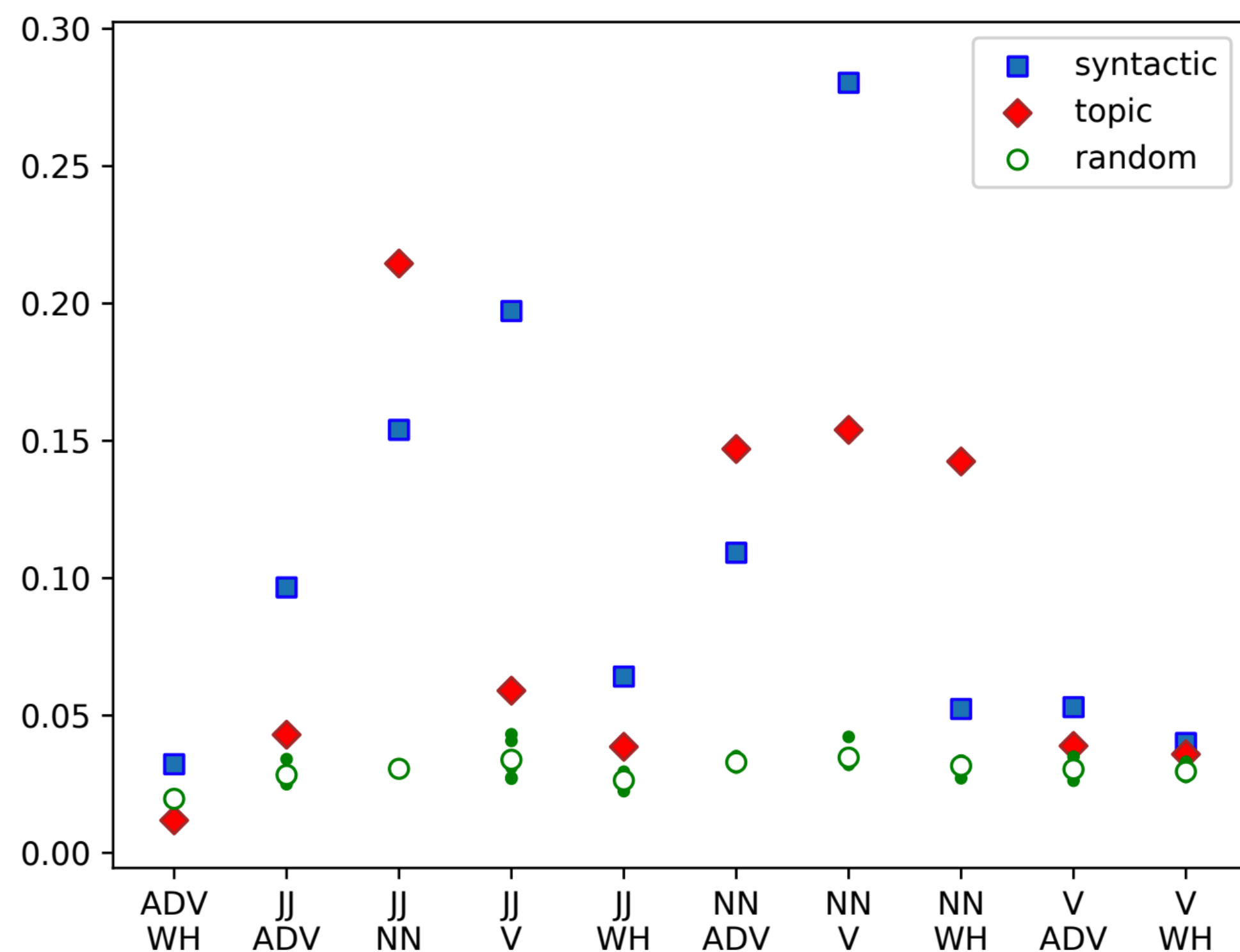
※ これを解くアルゴリズム
云々は future work

実験：品詞タグによる分割から他の分割を識別できるか？

- いろいろな分割を作って、先の量に基づいて比較してみる
- データ：Simple English Wikipediaから構築
 - 各トピック430文、
合計語彙数2,963トークン
- 3トピック：Numbers, Democracy, Hurricane
- 言語 L は3-gramの集合（重複あり）、文の確率は3-gram頻度による
 - 長距離依存が入ってこないように
Simple Wikipediaを使ってる？
- P_{con} ：トピックごとの出現頻度による分割
 - 論文によると $I_L(P_{\text{con}}) = 0.06111$, $I_{\bar{L}}(P_{\text{con}}) = 0.06108$ でかなり文脈的
- P_{syn} ：Stanford taggerで自動付与した品詞の頻度による分割

細かく分割すれば当然情報量は増える

…が真に統語的な情報はどれか？



- いろいろな品詞ペアを統合した分割 P_{merge} から情報量の増加をプロット

- P_{syntax} : 品詞の組みをマージしない場合

- P_{topic} : マージした後トピックに基づき再分割
NNとVをマージ -> Numbers, Democracy, Hurricaneで再分割

- P_{random} : マージした後ランダムに2分割

Figure 3: Increase of information Δ_I in three scenarios: syntactic split, topic split and random split.

品詞タグによる真の統語的分割を識別できるか

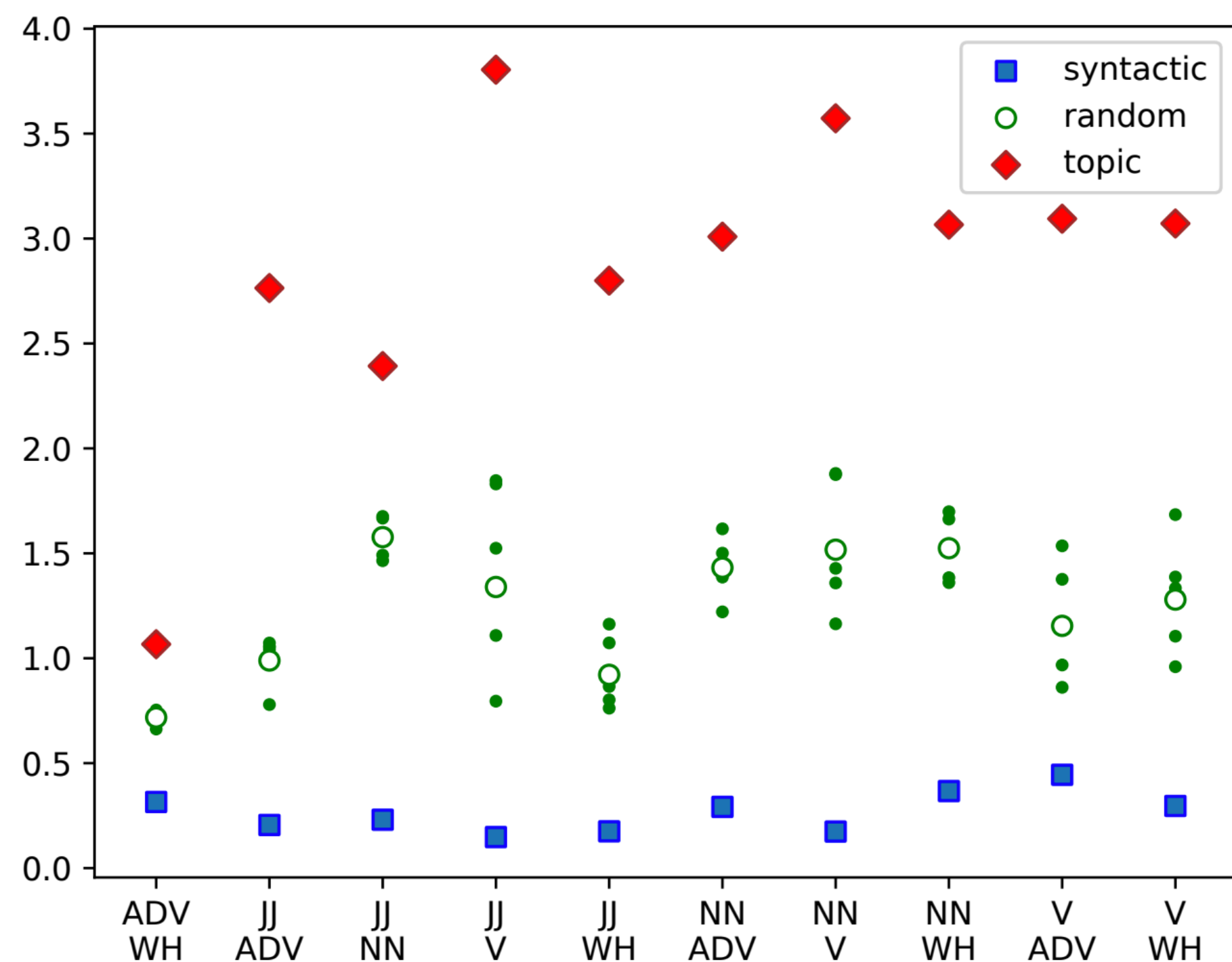


Figure 4: Ratio Δ_{MI}/Δ_I in three scenarios: syntactic split, topic split and random split. Considering objective (2) with parameter $\mu = 0.5$ leads to discrimination between contextual and syntactic information.

- P_{con} を使って $\min_P I_L(P; P_{\text{con}}) - \mu I_L(P)$ を考える
- 文脈的分割の代表 μ, γ 統語的: $\min_P \max_{P^*} I_L(P; P^*) - \mu I_L(P)$
- P, P' について以下が成り立てば P のほうが統語的
- $\frac{I_L(P; P_{\text{con}}) - \mu I_L(P)}{I_L(P) - \mu I_L(P')} \leq \frac{I_L(P'; P_{\text{con}}) - \mu I_L(P')}{I_L(P') - \mu I_L(P')}$
 $\leftarrow I_L^\mu(P)$ とする
- 書き換えて $\frac{I_L(P; P_{\text{con}}) - I_L(P'; P_{\text{con}})}{I_L(P) - \mu I_L(P')} \leq \mu$
 $\leftarrow \Delta_{MI}$ (numerator), $\leftarrow \Delta_I$ (denominator)
- 図は $P = \text{syn, random, topic}, P' = \text{merge}$ のプロット
- 図によれば $\mu = 0.5$ で全ての品詞対に対し、

$$I_L^{0.5}(P_{\text{syn}}) \leq I_L^{0.5}(P_{\text{merge}}) \leq I_L^{0.5}(P_{\text{random}}), I_L^{0.5}(P_{\text{topic}})$$

品詞による分割を他と識別できた！

まとめ/感想

- ・ 言語モデルから品詞タグのような構造を抽出する手法を提案
 - ・ 鍵は統語的/文脈的分割の情報理論による形式化
 - ・ 統語的分割の概念が直感に沿うような性質を持つかももう少し書いてほしかった
- ・ 3-gram言語においてアイデアの有効性を実証
 - ・ 複雑な言語でも期待通りの結果がみられるか？
- ・ 実際に言語モデルと組み合わせたりする話は今後の課題
 - ・ 分割の空間は凸っぽいのでなんとかできそう？
 - ・ Conclusion曰く 著者は今ここに取り組んでるそう

ここまでのおさらい

- ・ (確率的) 分割 $P = (C, \{\pi_v\}_{v \in V})$

分割による像、カテゴリ言語 $\pi(L)$

$$p_{\pi(L)}(c_1 c_2 c_3) = \sum_{v_1, v_2, v_3} \pi_{v_1}(c_1) \pi_{v_2}(c_2) \pi_{v_3}(c_3) p_L(v_1 v_2 v_3)$$

- ・ 分割 P が **文脈的** $\stackrel{\text{def}}{\Leftrightarrow} \pi(L) = \pi(\bar{L})$

- ・ 語順を shuffle して得た \bar{L} と分割による像言語が同じ構造

- ・ 分割 P が **統語的** $\stackrel{\text{def}}{\Leftrightarrow} P$ が任意の文脈的分割 P' に独立

- ・ 独立性 : $P, P', P \cdot P'$ の像言語にて $p_{\pi \cdot \pi'(L)}((c_1, c'_1)(c_2, c'_2)(c_3, c'_3)) = p_{\pi(L)}(c_1 c_2 c_3) p_{\pi'(L)}(c'_1 c'_2 c'_3)$