Pattern Lattice を使ったヒトの言語知識と処理のモデル化

黒田 航* & 長谷部 陽一郎**

*NICT　　**同志社大学/NICT

概要

- Pattern Lattice の理論
- Pattern Lattice 処理系の実装例の紹介

Theory of Pattern Lattice

出発点

- ヒトは文 $s = w_1\ w_2 \ldots w_n$ の意味を，s を構成している語 $w_1, w_2, \ldots, w_n$ の語彙的意味を合成して得ているとは考えにくい
  - そうだとしたら，自然言語の文の意味はもっともっと規則的であるはず
  - 機械翻訳はもっともっとうまくいってよい
- 自然言語の意味が規則的/構成的というのは幻想

非構成性の簡単な例 1/2

- 次の二つの文のでは同じ動詞「かかる」が使われているのに，多くの人は意味が似ているとすら感じない

  (1) その男は医者にかかっていた

  (2) その絵は壁にかかっていた

- しかし ...

非構成性の簡単な例 2/2

- 次のような例で生じる語義の競合は説明不能

  (3)??その絵は医者にかかっていた

  (4)??その男は壁にかかっていた

- 「かかる」の語義の曖昧性だけで説明できるか?
  - ムリではないかも知れないが効率は悪い
  - (1, 2)の例で同じ効果が起きない理由が説明できない

見こみのある路線 1/2

- 次のような超語彙的パターンに(語の意味に還元不可能な) 状況喚起の効果を認めるのが効率的

(5) X_1 は壁にかかっていた

- X_1の典型的な実現値は {(その)絵, (その) 服, (その) コート, (その) 帽子, ...}

(6) X_2 は医者にかかっていた

- X_2の典型的な実現値は {その人, 彼, 彼女, (その)男, (その) 少年, ...}

発表後の補足

- (5, 6) の他に次のような超語彙的パターンの影響もある

(7) その絵は X_3 にかかっていた

- X_3 の典型的な実現値は {壁, 廊下, 玄関, 居間, ...}

(8) その男は X_4 にかかっていた

- X_4の典型的な実現値は {病気, 医者, X医, 病院, ...}

見こみのある路線 2/2

- 自然言語の意味は (5, 6, 7, 8) のような超語彙的パターンからの誘引で決まる
  - 慣用句やコロケーションは超語彙的パターンの特殊な場合
  - 超語彙的パターン非線型表現
- それらで決まっていない “隙間” の部分が語の意味で “埋め” られる

本発表の立場

- 新たな問題
  1. 超語彙的パターンはどれぐらい存在するか?
  2. 意味構築が構成的でないなら，どうやって新奇な表現の意味が理解できるのか?
- 膨大な事例記憶の上の Pattern Lattice の下での処理を考えることで，これらの問いに同時に答える

データ観察から

- 規模の大きなコーパスを調査しても，完全に同一な文が現われる可能性はかなり低い
- その一方で，ほとんどの表現が数百個程度の基本的なパターンの変異形 (variations)
  - 多くの表現にも複数個のパターンが同時並行的に部分一致する
- ただし

問題1への解答

- ヒトの言語知識が膨大な事例記憶 (黒田 2007, Port 2007)に基づくものであれば，超語彙的パターンは次の形で (原理的には) 際限なく存在する
  - 基本形の変異 (= 1次変異)
  - 変異形の変異 (= 2次変異)
  - 変異形の変異形の ... の変異 (= n次変異)
- Pattern Lattice はこの問題を合理的に解決

問題2への解答 1/3

- 非構成的意味構築のモデル化の具体案
  - 任意の表現 e について，e に同時並行的に部分一致するパターン群 $p_1, p_2, ..., p_n$ の間にうまく統語/意味演算を定義すれば，アナロジーに基づいた統語/意味処理の問題は解決する
  - 演算は $p_1, p_2, ..., p_n$ の素性の重合わせ (論理和)で ok
- これは (Parallel) Pattern Matching Analysis: PMA (Kuroda 2000; Kuroda & Iida 2005)の基本的発想

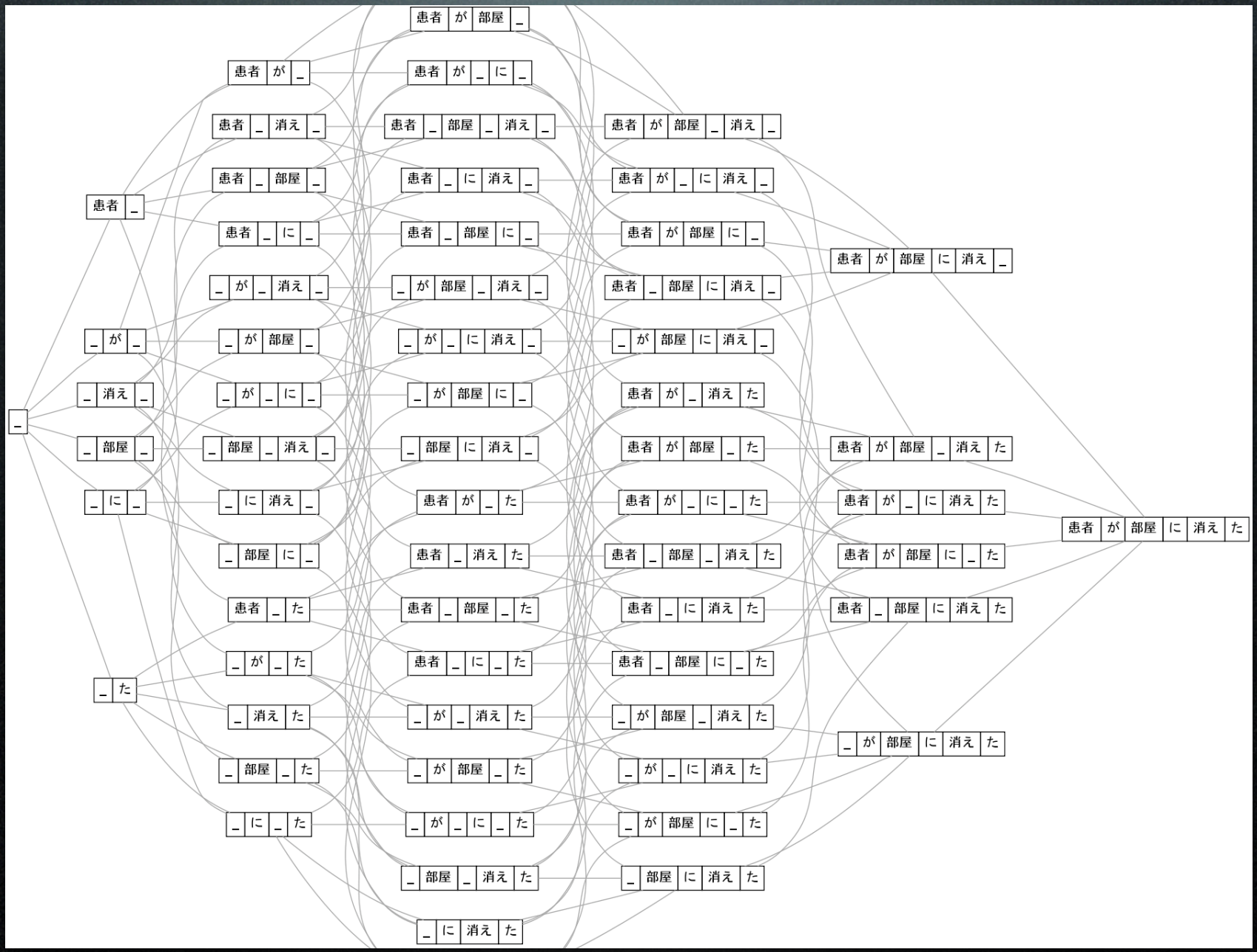
問題2への解答 2/3

- PMA によるモデル化の難点は，$p_1, p_2, \ldots, p_n$ を網羅的に列挙するアルゴリズムが不在だった点
- その不備を補うのが本発表の Pattern Lattice の理論

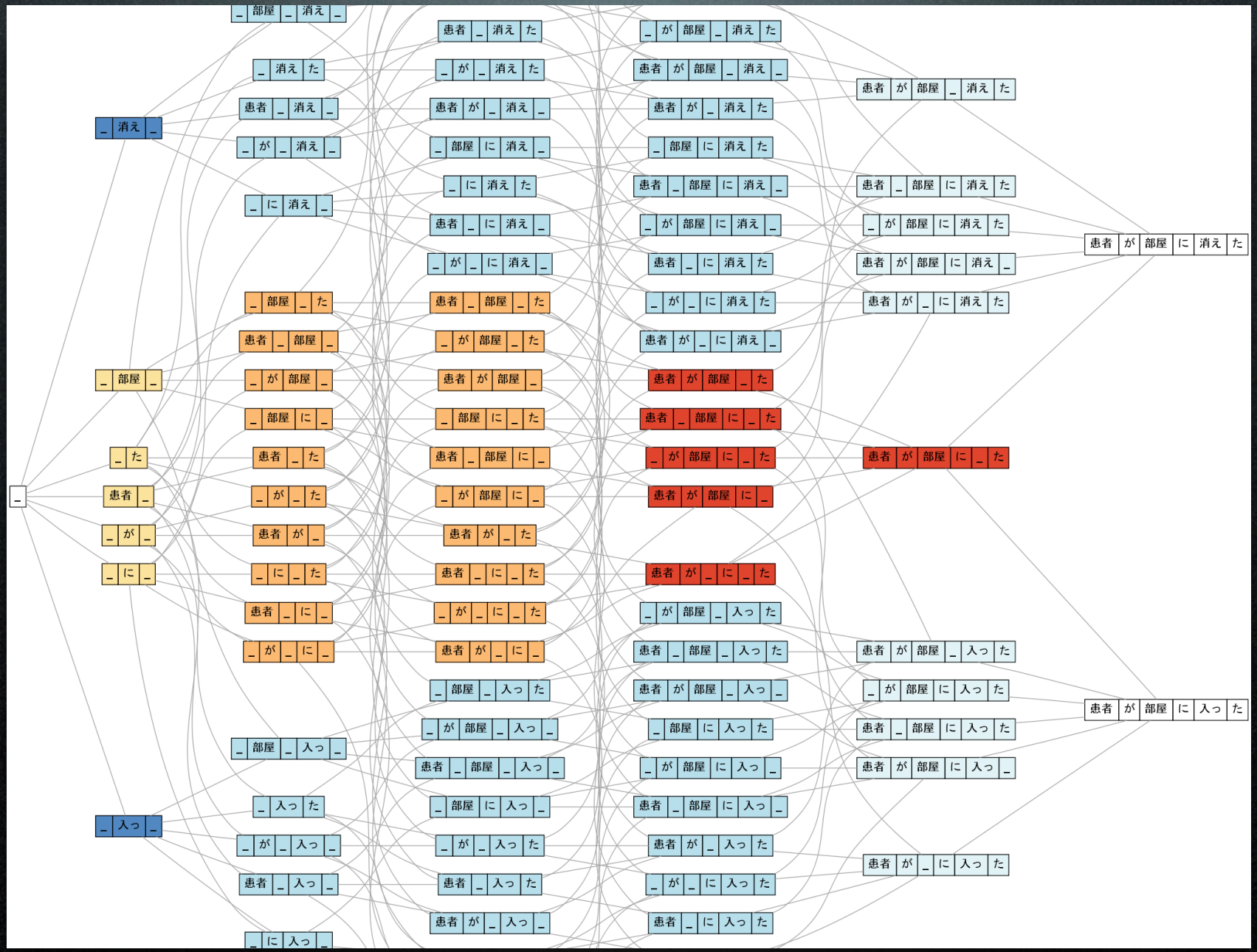
Pattern Lattice in Action

生成アルゴリズム

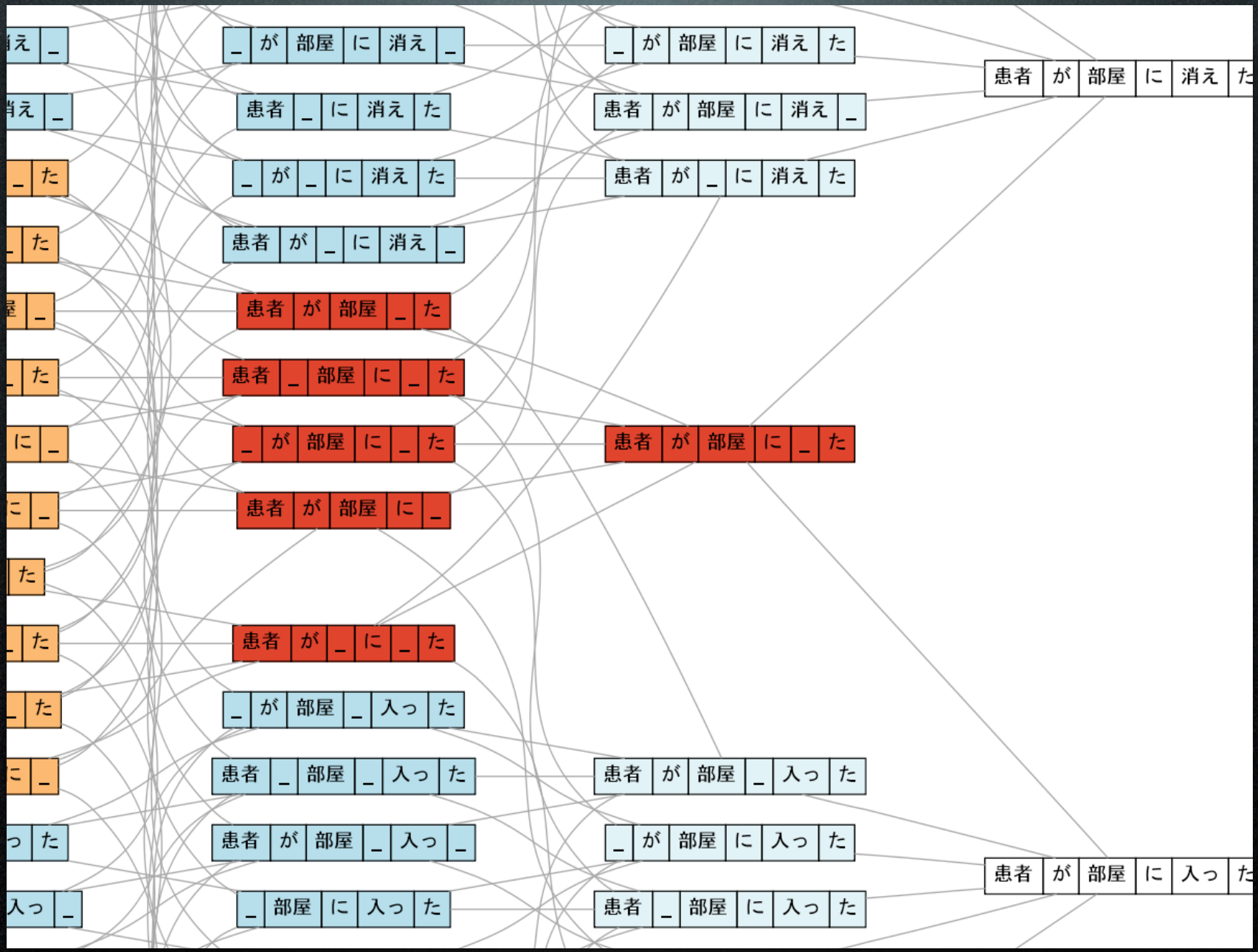
1. 表現 e を適当な単位 u_1 u_2 ... u_n に分割する
2. u_i を再帰的に変項化 (変項化の結果 = pattern)
3. 2で生成されたパターン集合の instance-of の下での半順序集合が Pattern Lattice
4. instance-of(p_i, p_j)の定義:
    1. is ($p_{i,k}$, $p_{j,k}$) OR
    2. instance-of ($p_{i,k}$, $p_{j,k}$)


患者 が 部屋 _
患者 が _
患者 が _ に _
患者 _ 消え _
患者 _ 部屋 _ 消え _
患者 が 部屋 _ 消え _
患者 _ 部屋 _
患者 _ に 消え _
患者 が _ に 消え _
患者 _
患者 _ に _
患者 _ 部屋 に _
患者 が 部屋 に _
患者 が 部屋 に 消え _
_ が _ 消え _
_ が 部屋 _ 消え _
患者 _ 部屋 に 消え _
_ が _
_ が 部屋 _
_ が _ に 消え _
_ が 部屋 に 消え _
_ 消え _
_ が _ に _
_ が 部屋 に _
患者 が _ 消え た
_
_ 部屋 _
_ 部屋 _ 消え _
_ 部屋 に 消え _
患者 が 部屋 _ た
患者 が 部屋 _ 消え た
_ に _
_ に 消え _
患者 が _ た
患者 が _ に _ た
患者 が _ に 消え た
患者 が 部屋 に 消え た
_ 部屋 に _
患者 _ 消え た
患者 _ 部屋 _ 消え た
患者 が 部屋 に _ た
患者 _ た
患者 _ 部屋 _ た
患者 _ に 消え た
患者 _ 部屋 に 消え た
_ が _ た
患者 _ に _ た
患者 _ 部屋 に _ た
_ た
_ 消え た
_ が _ 消え た
_ が 部屋 _ 消え た
_ が 部屋 に 消え た
_ 部屋 _ た
_ が 部屋 _ た
_ が _ に 消え た
_ に _ た
_ が _ に _ た
_ が 部屋 に _ た
_ 部屋 _ 消え た
_ 部屋 に 消え た
_ に 消え た

[患者,が,部屋,に,消え,た]



[患者,が,部屋,に,{消え,入っ},た]



[患者,が,部屋,に,{消え,入っ},た]

Pattern Lattice Builder

- www.kotonoba.net/rubyfca/pattern
  - 分割数が 6 を超えると Graphviz/dot の処理が重くなるので注意されたし

注意

- PL上の処理では単語の合成ではなく，超語彙的パターンの合成によって目的を実現するが
- 意味のある部分の合成によって全体を構成する点では従来のモデルと本質は変わらない

Pattern Lattice の問題点

- 扱える要素の数に上限がある
  - 要素の数が7を超えた辺りから急に s/n が大きくなる
  - 複数のレベルで Pattern Lattice が成立するのでは?
- 記憶容量より検索の効率化が問題
  - 並列処理を想定しても効果的な索引づけが必要
  - ヒトの想起の仕組みにトリックがあるのでは?

注意 1/2

- 表現の分割が任意なのは意図的
  - 音素の集合=>形態素
  - 形態素の集合=>語
  - 語の集合=>句
  - 句の集合=>文
- のような厳密にボトムアップな構成系を考えているわけではない

Summary

発表のまとめ

- 自然言語の意味の非構成性を捉えるためにヒトの言語知識を Pattern Lattice としてモデル化
- 語彙意味論で説明のつかない現象の説明の可能性を提示し，
- 試験的な実装を紹介した

今後の課題

- 大規模化/データベース火の可能性を検討したい
  - 今は使い捨てだが，できれば処理結果をデータベースとして蓄積する仕組みを導入したい
- パターンを素性表現して階層性を暗黙化したい
- 変項を意味クラスとして特徴づける仕組みを導入したい
  - 今は文字列一致しか扱えていない

Thank you

Discussion

知性観

- ヒトが知的なのは
  - すぐれた知性を備えているからというより
  - 膨大な事例記憶を効率良く使っているから
- 関連する議論
  - Hawkins (2004) の Memory-Prediction framework

記憶という概念の明確化 1/2

- 覚え (storage) と思い出し=想起 (recall/ remembering)は別の処理
- 覚えには上限がないが，思い出しには強い制限がかかっている
  - 更に言うと想起の基本的仕組みは検索 retrieve ではない

記憶という概念の明確化 2/2

- 動物の記憶には想起可能な記憶 explicit memory と想起不可能な記憶 implicit memory が共存
  - てんかんの治療で不幸にして後行性健忘者になった HM は explicit memory は失ったが implicit memory は失っていない

記憶のパラドックス

A. 知覚したことは覚えていない限り思い出せない

- 将来に必要になるかどうかを見越して覚えるか否かを先決できない => 盲目的に覚えるしかない
- 無用な想起は正しい現実認識の邪魔になる
- 患者 S の症状

B. 覚えたことの多くは必要がない限り思い出さない方が適応的

パラドックス解消の条件

1. 何から何まで全部覚える
2. 効率的な思い出しのための効果的なインデクスづけを行なう
   - 睡眠時の脳の活動の一部はこれ
3. しかし，実際の想起は思い出しに対する恒常的な抑制の一時的な弱化によって起こる
   - 月元 (2008) の EMILE モデル

Vast Memory の証拠

- Solomon Shereshevky (Luria のMnemonist)
- Kim Peek (Savant Syndrome)
- は通常のヒトとどう違うか?
- 彼らは異常な銘記能力を獲得したというより無用なことを想起さない能力を失っているだけでは?

結論の系 1/2

- カテゴリー事例記憶 exemplar memory モデル (Nosofsky 1993ほか) は正しい
  - ヒトの知性は膨大な事例記憶 Vast Exemplar Memory: VEM の上に成立している
- Case-based Reasoning システム (Kolodner 1993 ほか) は正しい

結論の系 2/2

- ヒトは自分が知覚したことありとあらゆることをそのまま覚えているが，そのほとんどが想起できない状態にある
- 言語の知識もそういう種類の膨大な事例記憶の上に成立していると考えると「文法」の役割は極力小さくできる
- それと同時に単語の辞書は意味をもたなくなる