# 言い換え技術の研究動向 (Mid 2018)

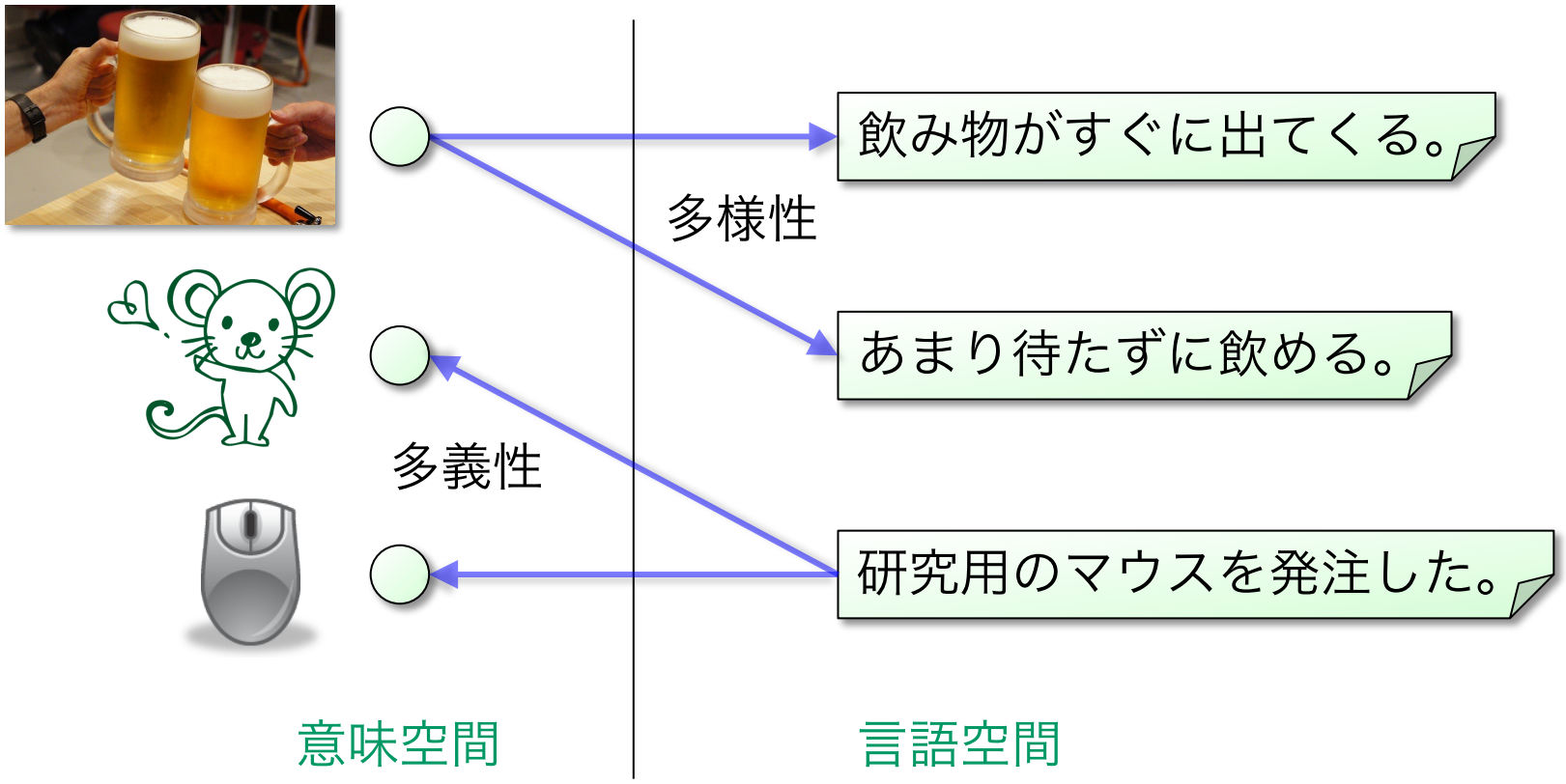
藤田 篤
情報通信研究機構 (NICT)
*http://paraphrasing.org/~fujita/*

# 本資料について

- テキスト量は多め
  - サーベイ資料として作成・更新
    - ただし数式や性能評価の数値はほとんど含まず
  - 講演に用いる際，すべてには言及しない
- 凡例
  - オブジェクト: 言語表現 知識の例
  - 色塗り文字(これ，これ):
    対応箇所，差分，ハイライト (要統一)
- 参考文献の見出しはWeb上の文献リストに対応
  - *http://paraphrasing.org/*
  - 最終更新は2015年9月 → 近々更新予定

# 言い換え (paraphrase)

- 換言，言い替え，パラフレーズ
  - モノ: 同じ意味内容を表す，同言語の異なる言語表現
  - コト: ある表現の言い換えを生成する行為



# 同じ意味を表す言語表現

- 言い換え: 同一言語内
  - 生成時: 必ずしも全体を変更するとは限らない，目的指向

出生届を居住地の市区町村の役所に提出してください。

お住まいの市区町村の役所に出生届をご提出ください。

- 翻訳: 異言語間
  - 生成時: 一部を除いて全体を変更，訳すこと自体が目的

お住まいの市区町村の役所に出生届をご提出ください。

Please submit notification of birth to the public office of the municipality where you live.

# 言い換えに関する研究の全体像

- 5つのレイヤ

**(L5) 応用技術への導入**
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

**(L4) 汎用的な言い換え計算機構**
言い換え認識，言い換え生成，評価・誤り分析

**(L3) 言い換えの計算に必要な知識の整備**
大規模知識獲得，言い換え知識ベース・コーパスの構築

**(L2) 事例研究**
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

**(L1) 分類体系**
定義と境界，現象の網羅・類型化，多言語対照分析

# 工学的な研究

- 基礎と応用

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(L4) 汎用的な言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

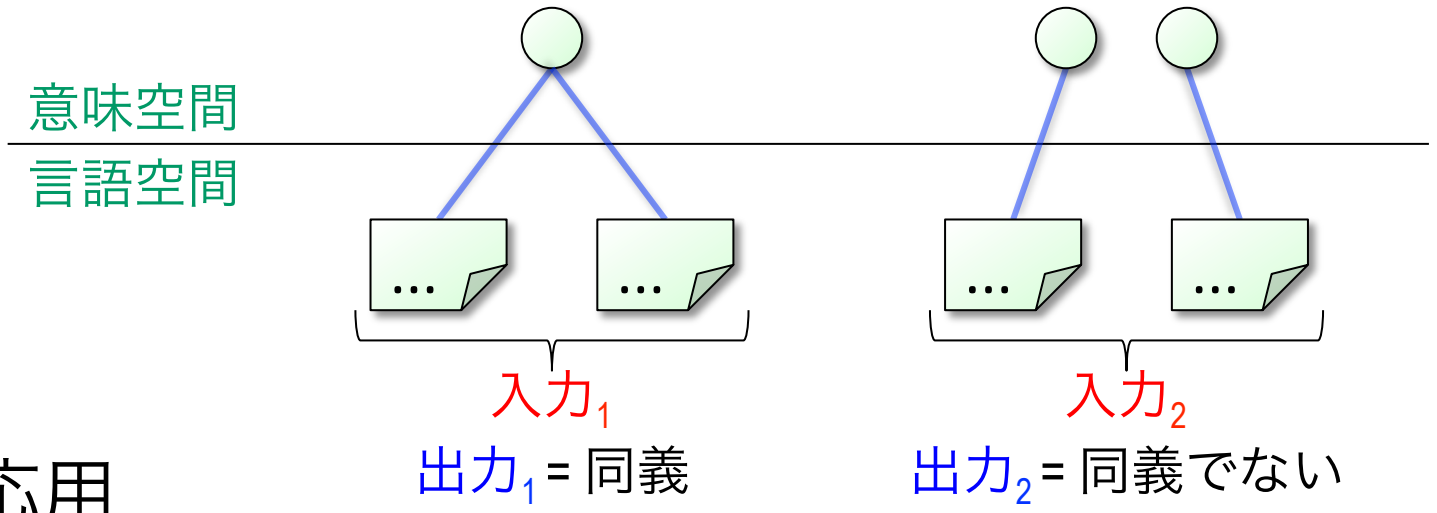
(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

# (L4) 言い換えの計算機構 (1/2)

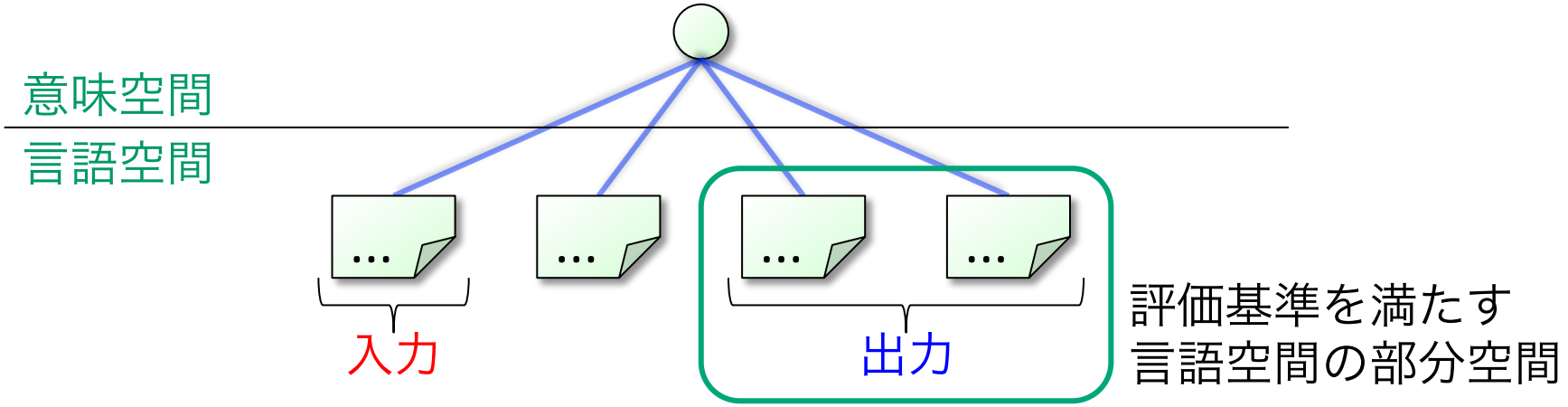
- 言い換え認識 (paraphrase recognition / identification)
  - 入力: 同一言語の複数(一般的に2つ)の異なる言語表現
  - 出力: 入力された言語表現が同義か否か(2値分類)
    - 似ている程度([0,1]や[1,5])を求めさせる場合もある




- 応用
  - 情報検索, 剽窃検出
  - 質問応答, 複数文書要約 (言い換え生成も併用)

# (L4) 言い換えの計算機構 (2/2)

- 言い換え生成 (paraphrase generation)
  - 入力: ある言語表現，目的に応じた評価基準
  - 出力: 入力と同じ意味，同じ言語の異なる言語表現の集合
    - ただし，与えられた評価基準を満たすもの




- 応用
  - 人間向け: 簡単化，文圧縮，音声合成(TTS)の前処理
  - 機械向け: MT前処理 (e.g., 未知語→既知語)

# (L3) 知識の整備

- 言い換えの処理に必要な知識
  - 語の素性/意味記述
    - e.g., 項構造/意味役割，動詞の使役・受身の可否
  - 表現間の関係に関する知識
    - e.g., 同義の語句の辞書 (言い換え知識)
      - 言い換え知識の自動獲得 (paraphrase acquisition)

語 窮地 ↔ 苦境　重傷 ↔ 大ケガ

句 脚光を浴びる ↔ 注目される　重傷を負う ↔ 大ケガをする

    - e.g., 派生語，反義語，特質構造，上位-下位，部分-全体
- 全貌はまだまだ不明
  - 言い換えの種類(L1/L2)と一緒に考える必要あり [藤田+ 04]
  - End-to-Endのアプローチでは今のところ無視されがち
    - 本当に不要かはまだ分からない

# (L5) 応用技術と言い換え

| | 言い換え生成 | 言い換え認識 |
|---|---|---|
| 人間向け | 文章執筆支援<br>文章読解支援<br>TTS前処理<br>MT後編集<br>ステガノグラフィ | 剽窃検出<br>翻訳メモリ検索<br>情報検索<br>質問応答<br>東口ボ |
| 機械処理向け | MTのモデル拡張<br>MT前編集 | データマイニング |

文書要約

# (比較的)言語学的な研究

- 全体像の俯瞰と特定の範囲の現象の理解

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(L4) 汎用的な言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

# (L1) 言い換えの分類体系

- これは言い換え?

健は悟から車を買った。
悟は健に車を売った。

主題/焦点が異なるが
明示的意味は等しい

暑すぎませんか?
窓を開けてくれませんか?

語用論的意味が等しい

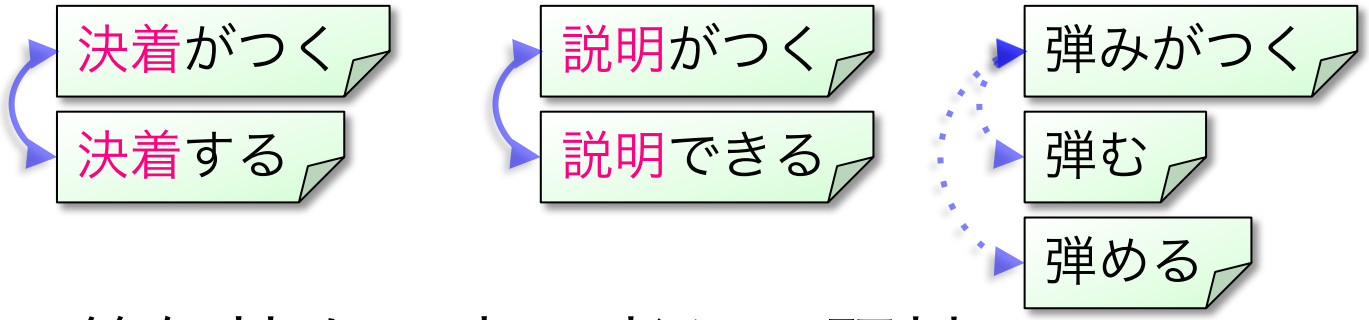
うちの子はトマトをよく食べる。
うちの子はトマトが好きだ。

常に真とは限らない

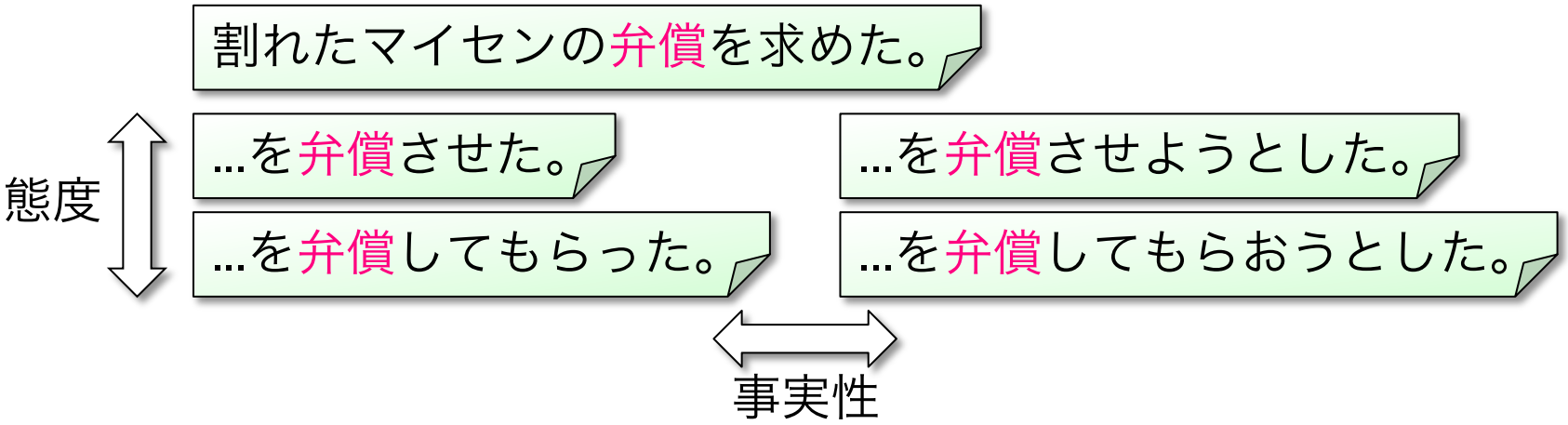
- 分類の意義
  - 人間が行っている計算の種類の俯瞰
  - 応用の際の取捨選択の判断材料

# (L2) 言い換えの事例研究

- 特定の範囲の言い換えが成立する原理の説明
  - e.g., 機能動詞構文と動詞句の言い換え
    - 「～がつく」




- 意味の等価性を丁寧に考える題材
  - e.g., 機能動詞構文と動詞句の言い換え



# 言い換えについて研究する意義

- コミュニケーションの円滑化
  - 中心的メディアは言語 (テキスト，音声)
    - 相手や状況に合わせて言い換える (言い換え生成)
    - 同じことを言っていると理解する (言い換え認識)
- 工学的重要性: 様々な応用の基盤技術
  - 個々の表現は必ずしも意味的に互いに素ではない
- 科学的興味: ことばの意味に迫るための良い例題
  - 本当に迫ろうとしているかどうかはさておき

# 本資料を作成した目的

- 研究の広がりや課題について知ってもらう
  - 工学的な研究
    - 3種類: 言い換え認識，言い換え生成，言い換え知識獲得
      - 深層学習の波
      - データセット
      - 評価における課題
    - 応用研究: 多数あるので簡単に
  - (比較的)言語学的な研究
    - 現象の網羅と類型化
    - 事例研究: 多数あるので簡単に
- (会話ができる)仲間を増やす
  - 地に足の着いた研究
  - データやツールを作ったり，使ってもらったり

# もくじ

(5) pp.85-91 (7)

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(2) pp.34-52 (19)　(3) pp.54-65 (12)

(L4) 汎用的言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(4) pp.67-83 (17)

(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

(1) pp.17-32 (16)　(6) pp.93-112 (20)

# 言い換えとは何か?

- 内包的定義: 辞書における語釈
  - e.g., 岩波国語辞典 (第7版)

> 言い換える: いったん言った言葉を他のに言い直す。
> また、(分からせるために)他の表現に変えて言う。

  - e.g., 新明解国語辞典 (第7版)

> 言い換える:
> (1)同じ内容を、別の言葉で説明する。
> (2)前とは別のことを言う。

  - e.g., 明鏡国語辞典 (第2版)

> 言い換える:
> (1)同じ事柄を別のことばで言う。
> (2)前言をとりやめて別のことを言う。

# 言い換えとは何か?

- 内包的定義: 辞書における語釈
  - e.g., LDOCE

paraphrase$_{(v)}$: to express in a shorter, clearer, or different way what someone has said or written

paraphrase$_{(n)}$: a statement that expresses in a shorter, clearer, or different way what someone has said or written

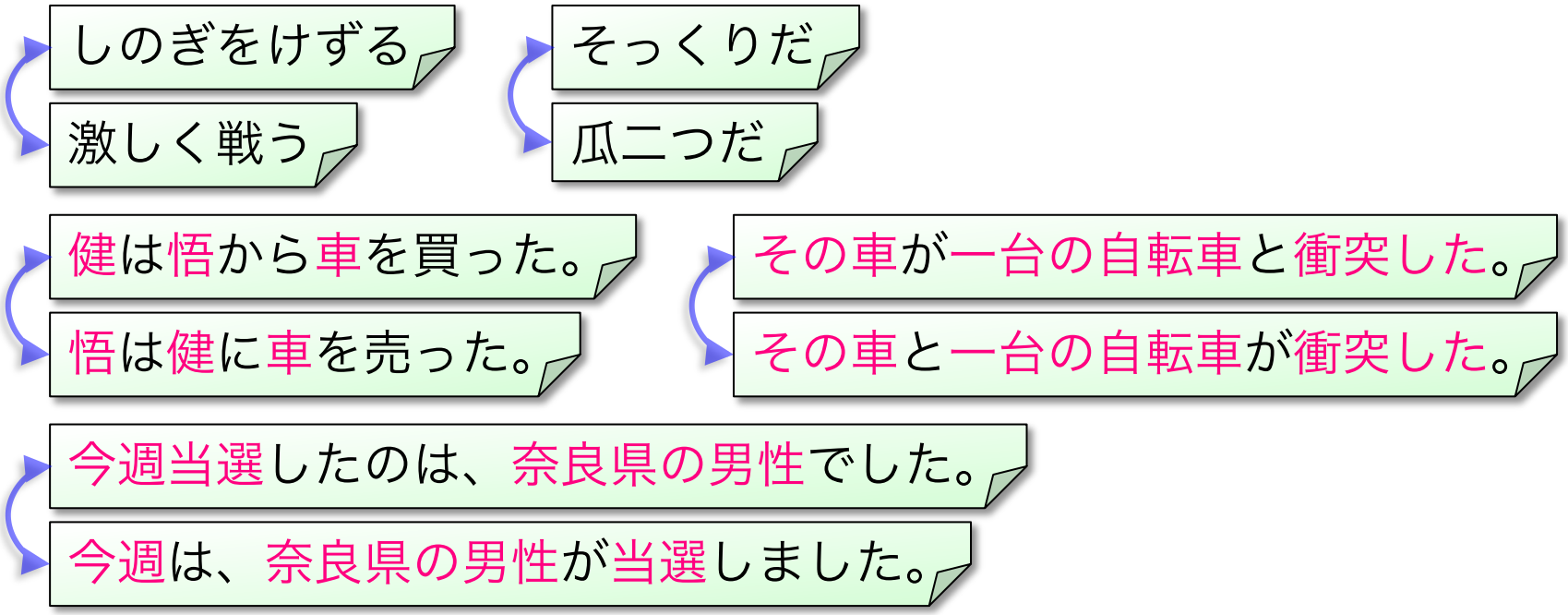
  - e.g., COBUILD

paraphrase$_{(v)}$: If you **paraphrase** someone or **paraphrase** something that they have said or written, you express what they have said or written in a different way.

paraphrase$_{(n)}$: A **paraphrase** of something written or spoken is the same thing expressed in a different way.

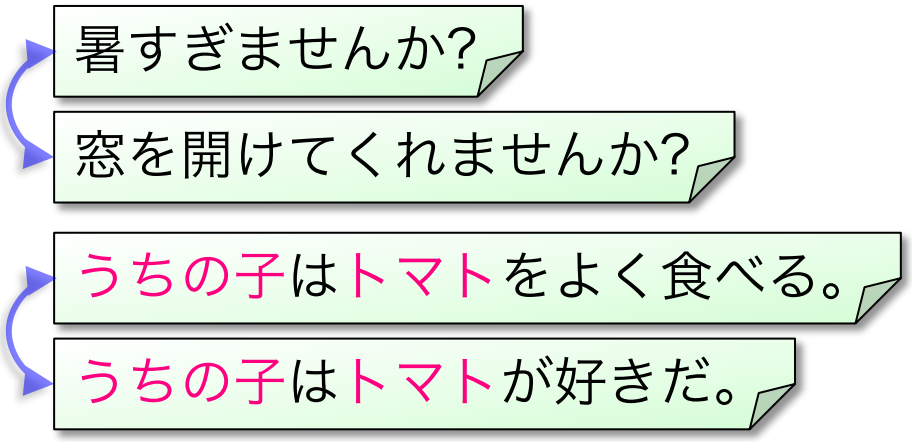
# 言い換えとは何か?

- 外延的定義: それらしいものを集めてみる
  - 語彙的, 構文的, 談話的, ...
  - 網羅できるか?



# 意味の同一性(同義性)に関する議論の意義

- 人間が行っている計算の種類の俯瞰
  - 全貌の解明，関連する概念との境界の整理
  - 「言い換えは大事ですよね〜」→ 本当に言い換え?




  - 自動化の際の道標
    - 目的や応用に応じた取捨選択，優先順位決め

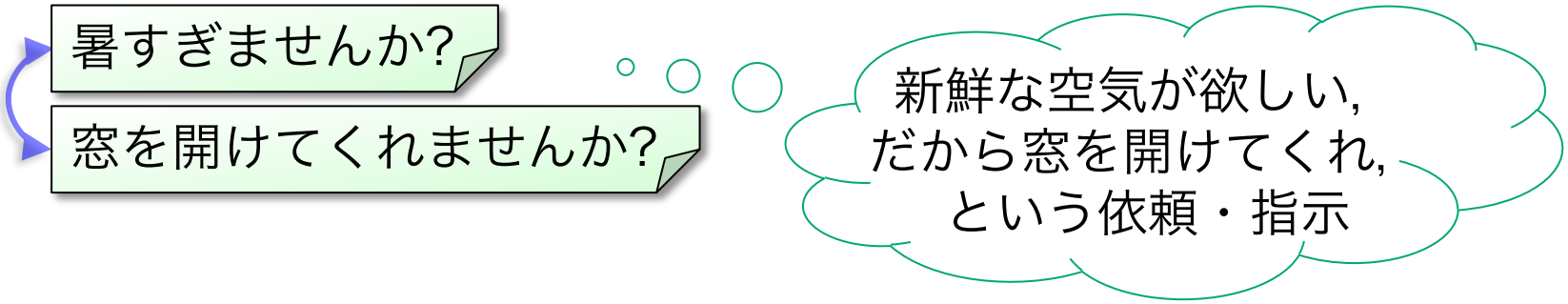
# 意味処理の例題としての言い換え

- 同義であるとは?
  - “Paraphrases are approximate conceptual equivalence among outwardly different material” [De Beaugrande+ 81]
  - “Every two forms (in language) contrast in meaning” [Clark 92]

- 同義性を考える3つのレベル [佐藤 99]
  - **a.** 語用論的意味が等しい
  - **b.** 参照先が等しい
  - **c.** 明示的意味が等しい

# a. 語用論的意味が等しい

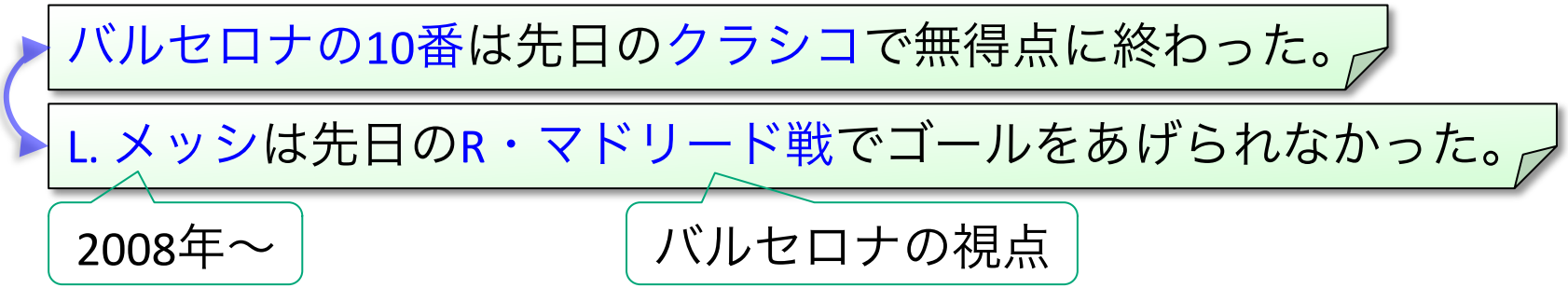
- コミュニケーションにおける発語の役割




- 言語行為論 [Austin 62]
  - 発語内行為 (Illocutionary act): 言葉の背後にある効力
  - 発語媒介行為 (perlocutionary act): 聞き手への効果
- 状況によって多様な解釈を持ちうる
  - 人間は推論しうる
  - ただし，真の解釈は話者のみぞ知る
    (匂わせておいて，そうではないとも言える)

# b. 参照先が等しい

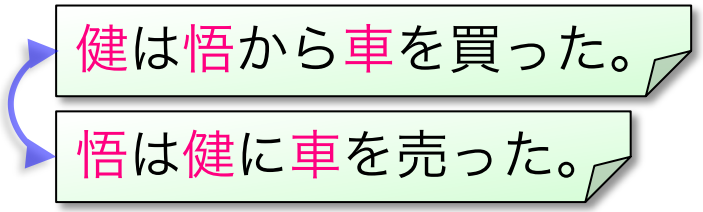
- 参照的意味 (referential meaning)




  - 異なる状況下では偽 (あるいは不自然)
    - e.g., ロナウジーニョ, リケルメ, リバウド, ...
    - e.g., バルセロナ戦, バルセロナとR・マドリードの試合
  - 談話あるいはそれを超えた意味
    - テキストの外側のエンティティの参照 (外界照応)
    - 文脈における認知的な意味 [Milićević, 07]

# c. 明示的意味が等しい

- 真理値意味論的意味 / 命題論理 / denotation




  - 誰でもが説明可能
    - 言語知識と世界知識に基づいて
    - 空気が読めなくても
  - 言外の意味 (connotation) は異なりうる [Edmonds 99][Inkpen+ 06]
    - 主題・焦点
    - 様式 (e.g., 形式性, 丁寧さ)
    - 感情・態度 (e.g., “スラっとした” vs. “ガリガリの”)

# 定義を振り返る

- 何らかの違いが想定・許容されている
  - e.g., LDOCE

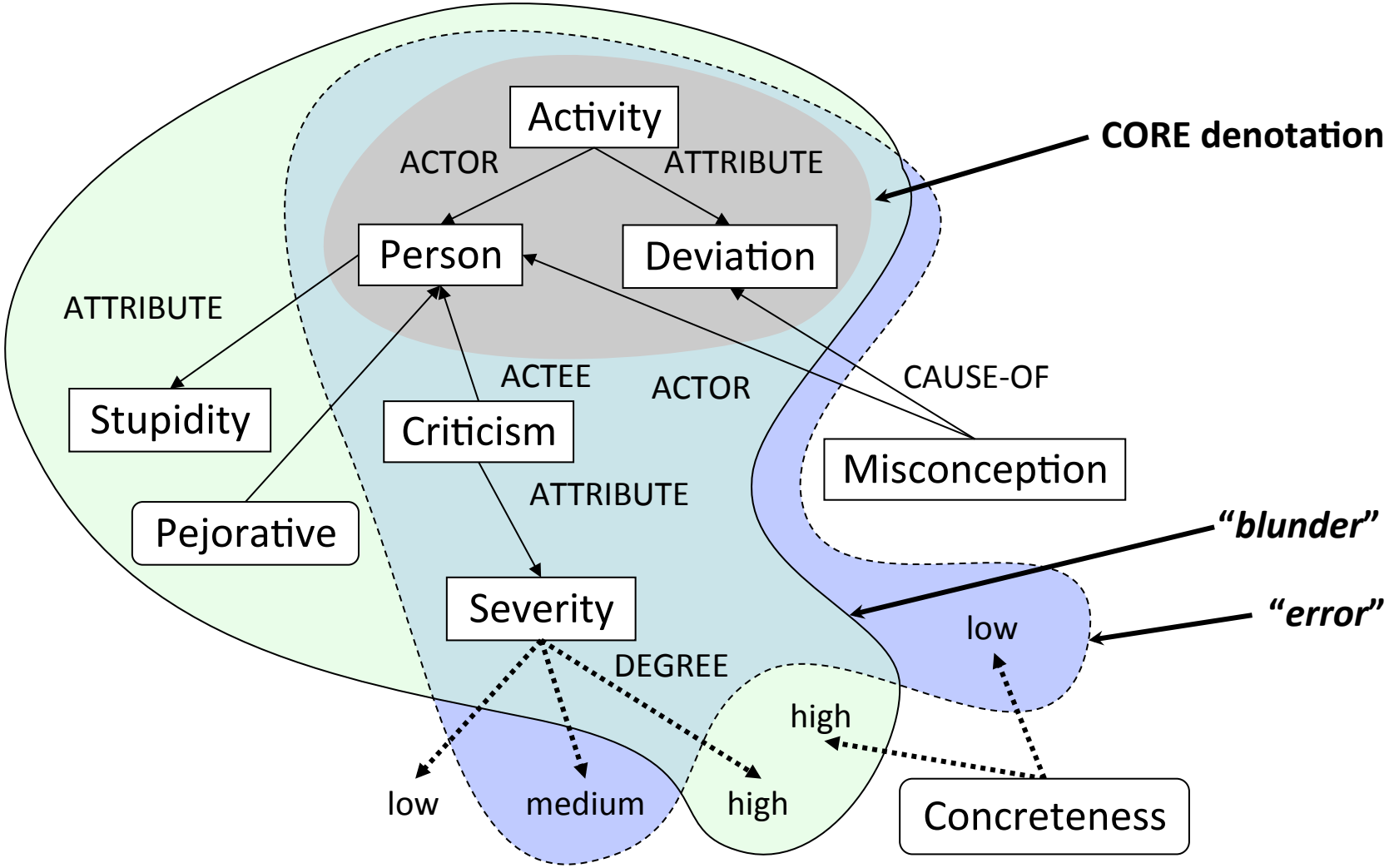
> paraphrase$_{(v)}$: to express in a shorter, clearer, or different way what someone has said or written

> paraphrase$_{(n)}$: a statement that expresses in a shorter, clearer, or different way what someone has said or written

  - “Every two forms (in language) contrast in meaning” [Clark 92]
    - Near-synonymous [Edmonds 99]
    - Approximate conceptual equivalence [De Beaugrande+ 81]
    - Quasi-paraphrases [Bhagat+ 13]
  - 特定の文脈において許容できる差を見極めるという問題

# “blunder” と “error” の比較

[Edmonds 99]

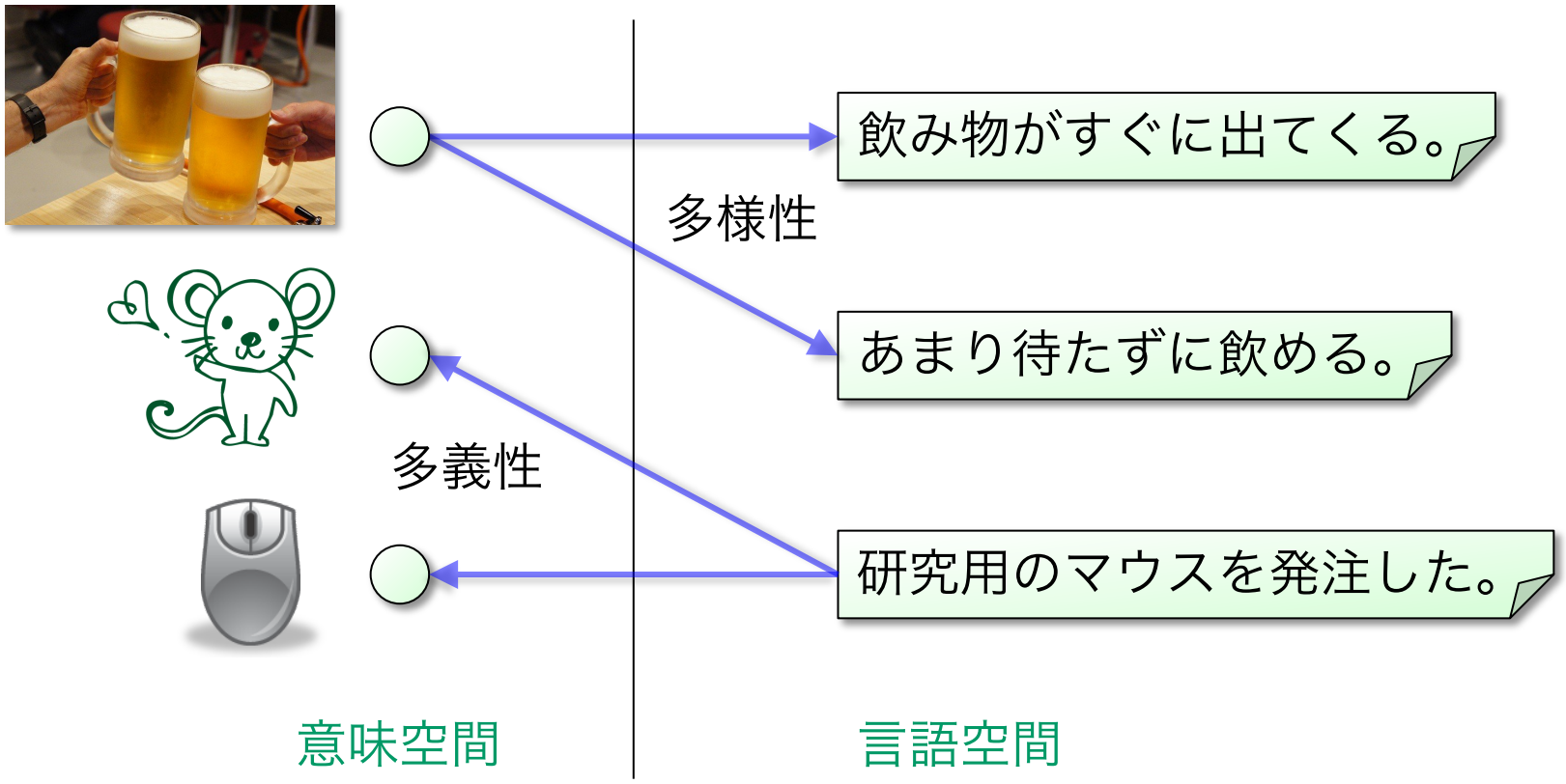


# 違いを生じさせるねらい

- 発信者と受信者の間の情報伝達の促進 (または阻害)
  - 人間どうしの対話
    - 発話内容の確認 [高塚 99]
    - 社会的な関係を保ったり押しつけたりするため
      - e.g., 仲間うちにおけるメンツや連帯感 [国広 00]
  - 目的に応じた要請
    - 簡単化: 難解 → 平易
      - 慇懃無礼化
    - 要約: 長い → 短い
    - 推敲: まわりくどい, ぎこちない, 冗長 → 明瞭
    - 客観化: 感情的 → 非感情的

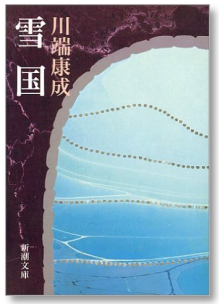
# 言い換え (paraphrase)

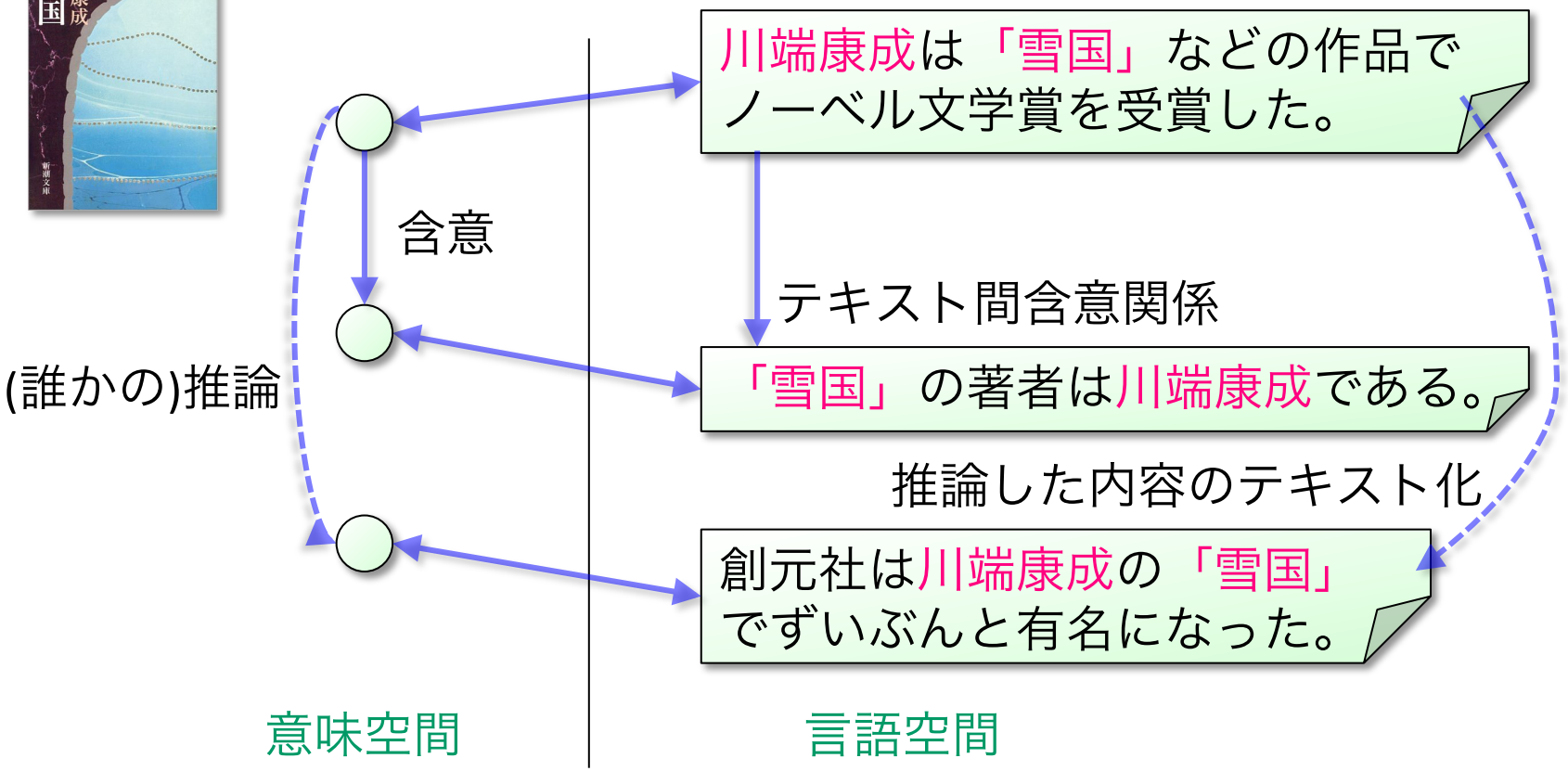
- 換言，言い替え，パラフレーズ
  - モノ: 同じ意味内容を表す，同言語の異なる言語表現
  - コト: ある表現の言い換えを生成する行為



# 関連する概念

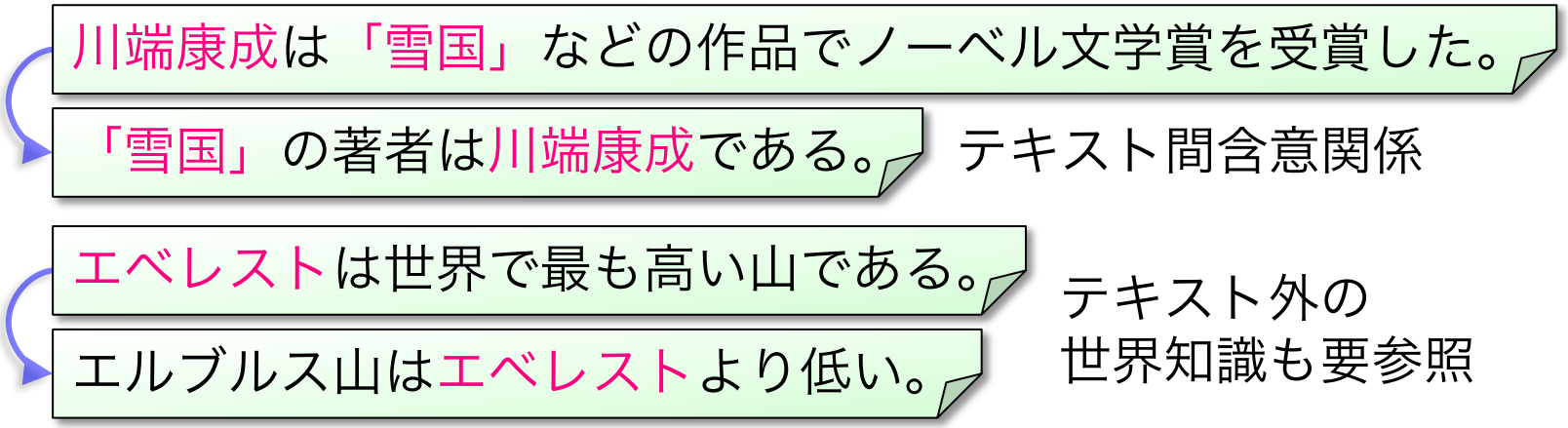
- 含意 (entailment), 推論 (inference)
  - 異なる明示的意味(それらに対応する表現)間の関係



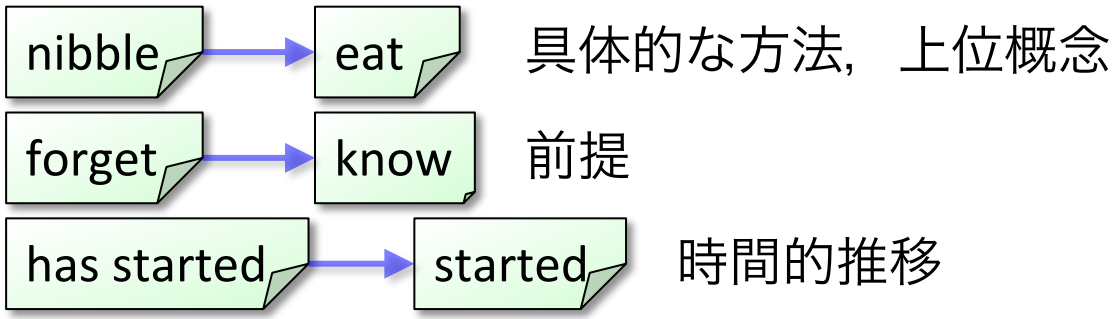



# 含意関係 (entailment)

- P → Q (Pが成り立つならばQが成り立つ)
  - Q→Pも真とは限らない




  - 語彙的含意 (lexical entailment) [Miller+ 85]




  - 言い換え ≡ 両方向の含意 [Dagan+ 05]

# (誰かの)推論 (inference)

- 真理値の同一性は保証されない
  - 数学や論理学における「導出」の意味の推論ではない
  - 誰もが同じように推論する/できるとは限らない
    - 語用論における発語内行為とも異なる
  - ただしこの計算が役立つ場合もある [Pantel+ 07]

川端康成は「雪国」などの作品でノーベル文学賞を受賞した。

創元社は川端康成の「雪国」でずいぶんと有名になった。

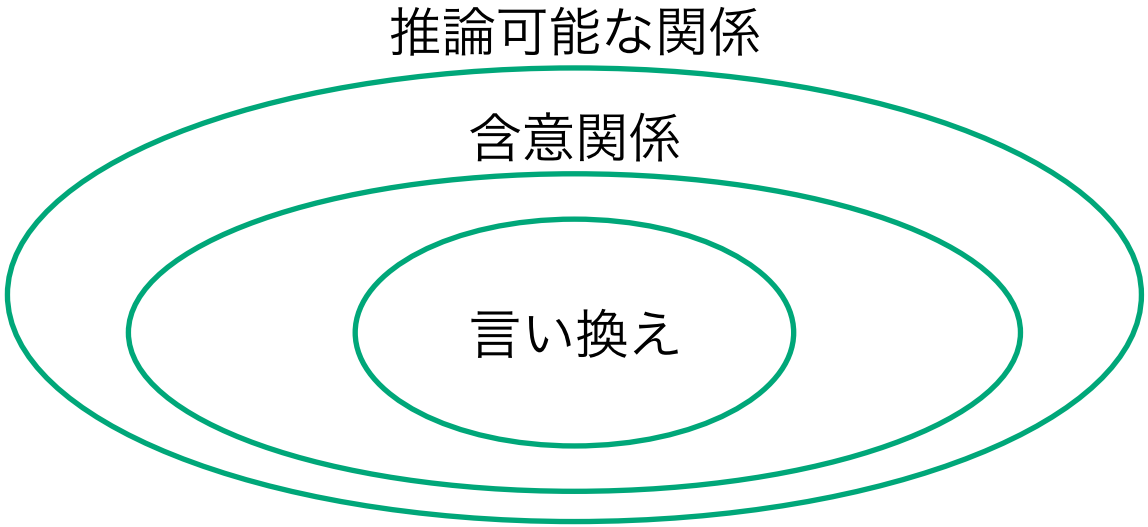
うちの子はトマトをよく食べる。

うちの子はトマトが好きだ。

# まとめ

- 同義性を考える3つのレベル [佐藤 99]
  - **a.** 語用論的意味が等しい
  - **b.** 参照先が等しい
  - **c.** 明示的意味が等しい

- 関連する概念



# もくじ

(5) pp.85-91 (7)

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(2) pp.34-52 (19)　(3) pp.54-65 (12)

(L4) 汎用的言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(4) pp.67-83 (17)

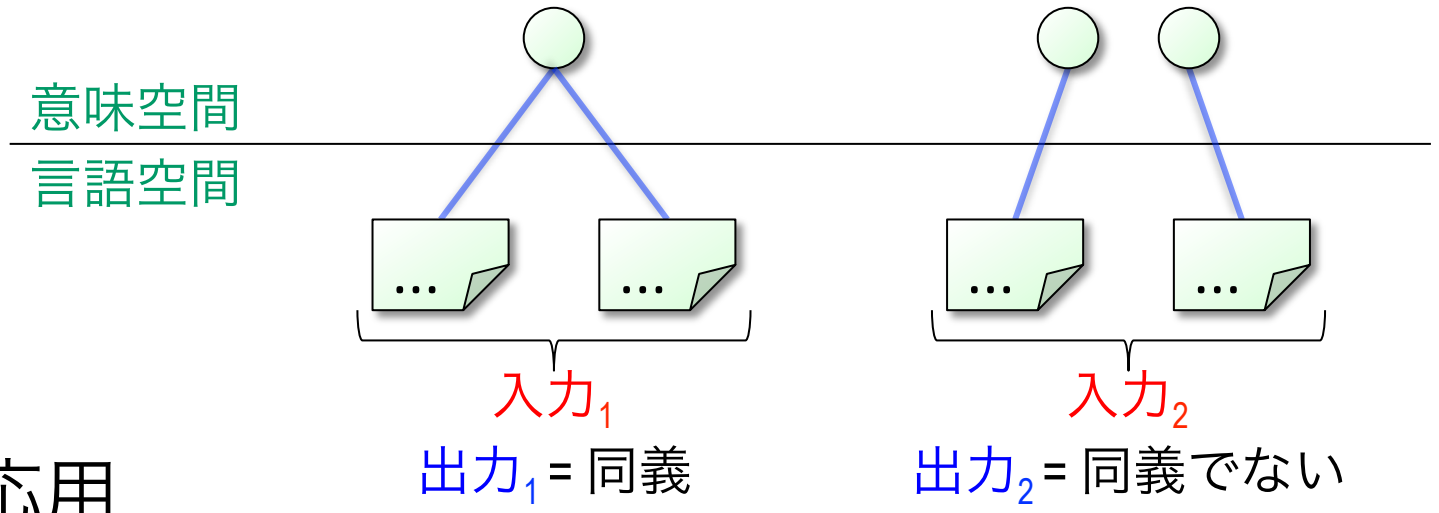
(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

(1) pp.17-32 (16)　(6) pp.93-112 (20)

# 言い換え認識

- タスクの定義
  - 入力: 同一言語の複数(一般的に2つ)の異なる言語表現
  - 出力: 入力された言語表現が同義か否か(2値分類)
    - 似ている程度[0,1]を求めさせる設定は適切か?




- 応用
  - 情報検索, 剽窃検出
  - 質問応答, 複数文書要約 (言い換え生成も併用)

# 2値分類問題/回帰問題としての定式化

- 教師あり機械学習の例題
  - 学習フェーズ
  - 判定フェーズ
- 性能評価
  - 学習データ同様の正解付きデータを用いる
  - 2値分類: 再現率，精度，それらのF値，AUC
    - STSでは相関係数(Pearson's $r$, Spearman's $\rho$)も用いられる
- 特殊ケース: 知識の自動獲得
  - 同義の文対の収集 [Dolan+ 04][Fader+ 13][Xu+ 14][Lan+ 17]
  - 同義の語句の収集 (= 言い換え知識獲得)

# 言い換え認識手法の大まかな分類

- シンボリックなアプローチ
  - 語句や部分構文木の対応関係の集約 [Takahashi+ 03][Finch+ 05][Madnani+ 12] [Watanabe+ 12]
  - 抽象概念上での照合 [Issa+ 18][Yanaka+ 18]

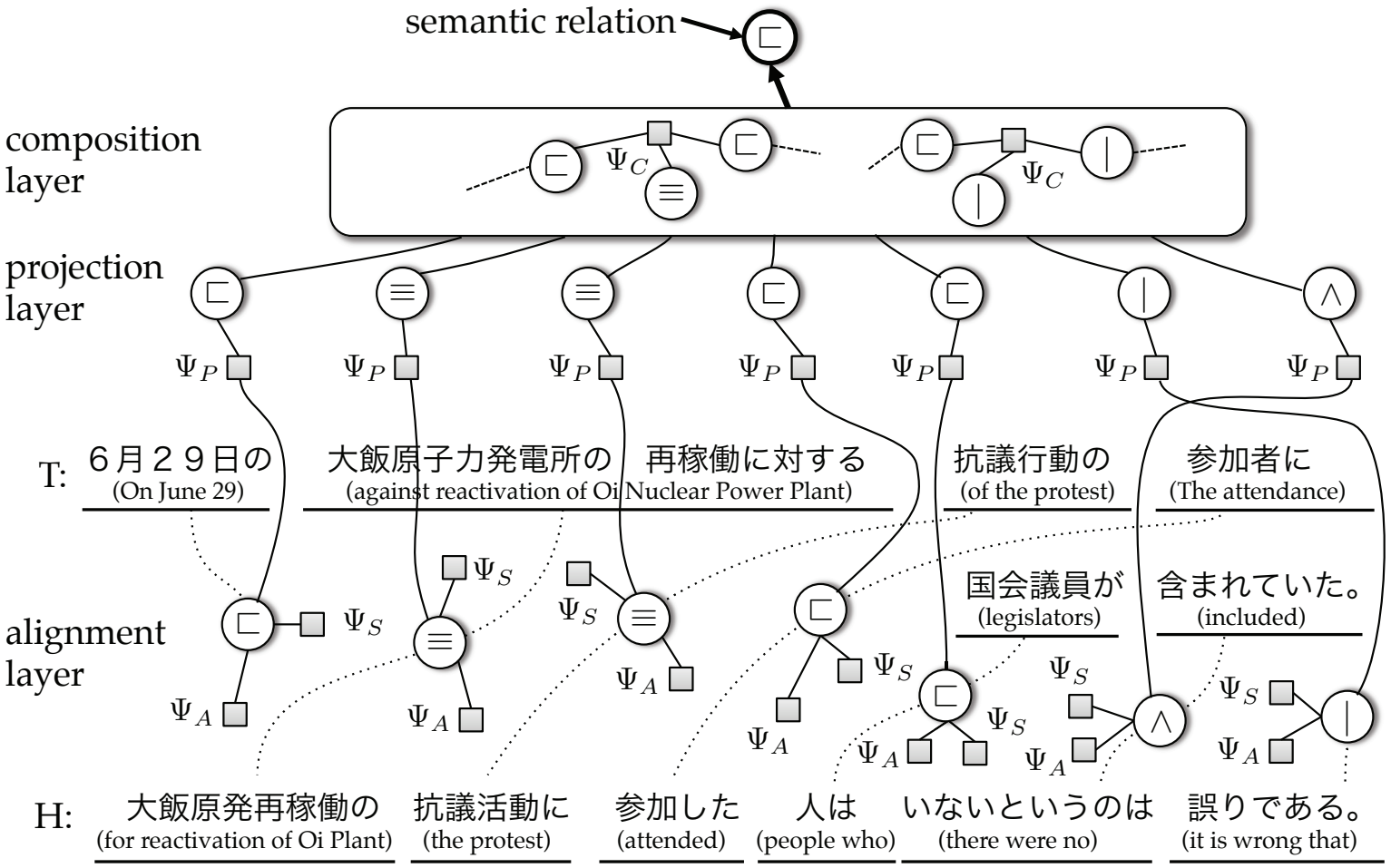
- 分散表現に基づくアプローチ
  - 文の分散表現の類似度(足し算最強説) [Wieting+ 16]
  - 語レベルの対応から総合的に判断 [Yin+ 15]
    - 複数のCNNフィルタと複数のプーリング [He+ 15]
    - 文脈を考慮した単語分散表現を用いた多層CNN [He+ 16]
  - 生成モデルによる文全体の変換尤度 [Mallinson+ 17]

# シンボリックなアプローチ (1/3)

- 素性
  - 表層的な素性
    - MT向け自動評価尺度: BLEU等 [Finch+ 05][Madnani+ 12]
    - 表層n-gramの重複度 [Liu+ 11]
  - 言語解析の結果得られる素性
    - 統語構造の照合結果 [Das+ 09]
    - 述語項構造の照合結果 [Watanabe+ 12]
    - Abstract Meaning Representation (AMR) の照合結果 [Issa+ 18]
- 定量化手法
  - SVMs [Finch+ 05][Das+ 09][Issa+ 18]
    - RBFカーネルがよく用いられる
  - 線形回帰 [Madnani+ 12]

# シンボリックなアプローチ (2/3)

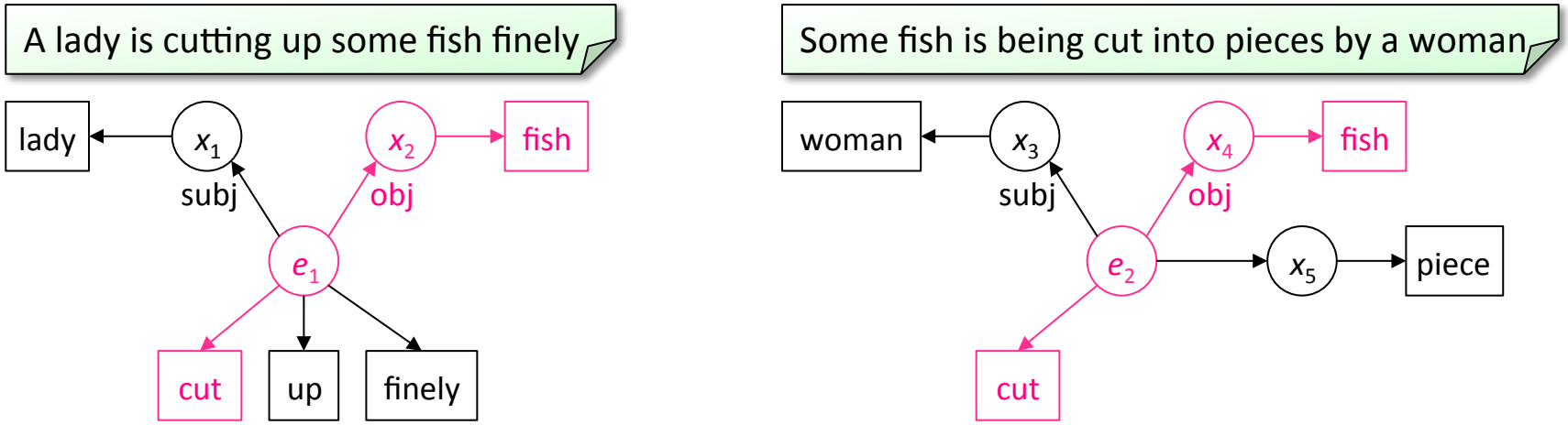
- 種々の語彙資源を用いた対応付け [Watanabe+ 12]




T: The attendance of the protest against reactivation of Oi Nuclear Power Plant on June 29 included legislators.
H: It is wrong that there were no people who attended the protest for reactivating Oi Nuclear Power Plant.

# シンボリックなアプローチ (3/3)

- 意味表現の照合
  - FrameNet [Ellsworth+ 07], AMR [Issa+ 18]
  - 基本論理式 [Yanaka+ 18]
    - 統語解析 → 高階論理式 → 基本論理式 (DAGで表現)
    - 学習時: 非対応箇所から公理を生成
    - 判定時: 学習済の公理を用いて非対応箇所を単一化



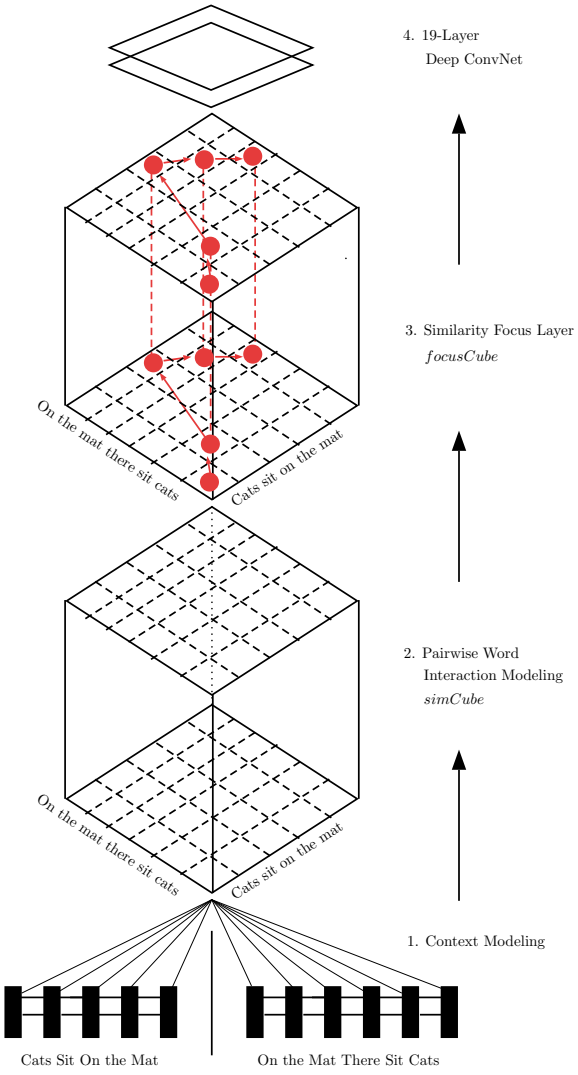

語 $\forall x_1$ (**lady**($x_1$) → **woman**($x_1$))

句 $\forall e_1$ (**cut**($e_1$)∧**up**($e_1$)∧**finely**($e_1$) → $\exists x_5$ (**into**($e_1$, $x_5$)∧**piece**($x_5$)))

# 分散表現に基づくアプローチ (1/3)

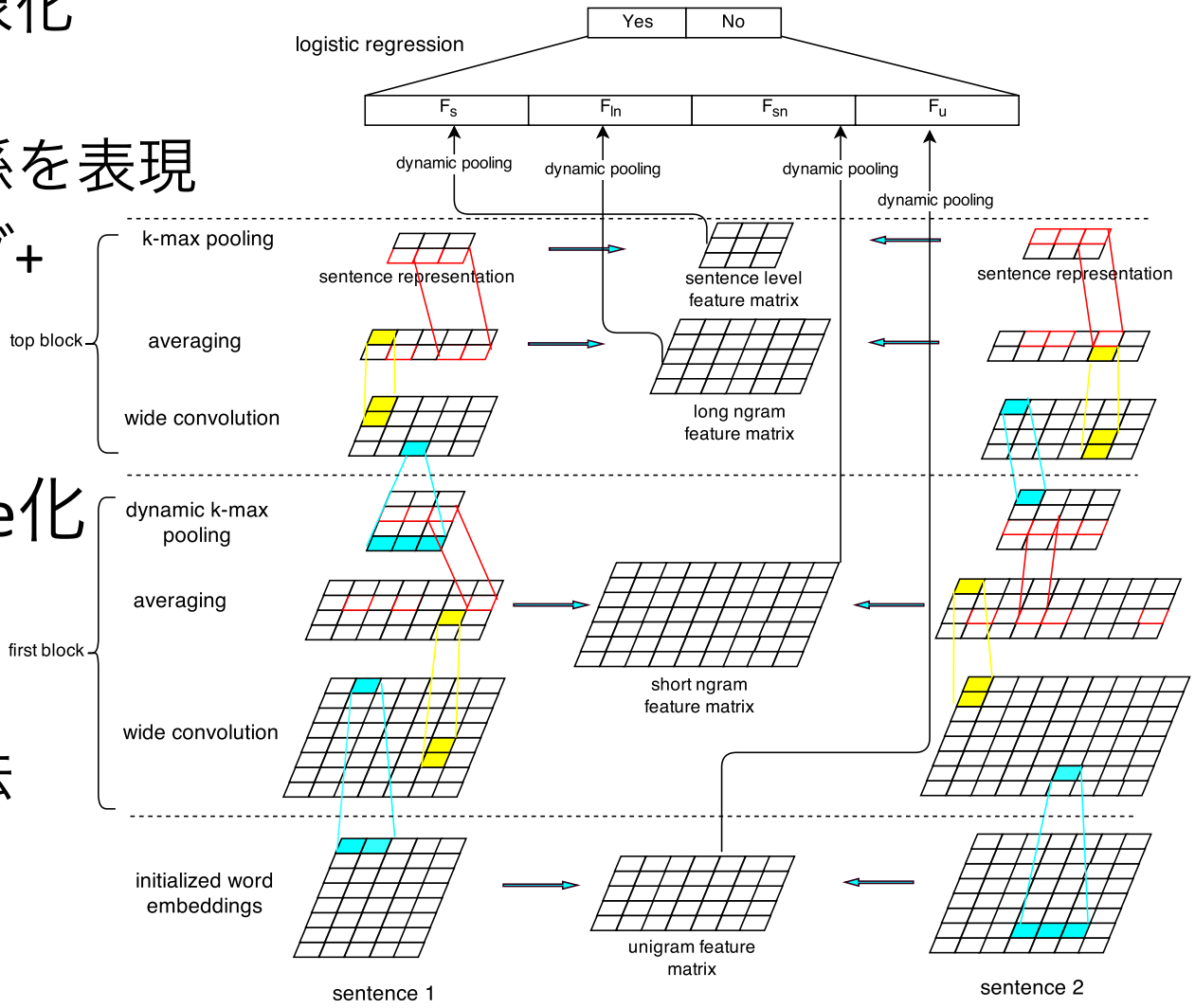
- 文の分散表現の類似度
  - 単語分散表現の平均 [Wieting+ 16]
  - LSTM [Wieting+ 16]
- 語レベルの対応から総合的に判断
  - 文中の各語の分散表現
    - 事前学習 [Mikolov+ 13]
    - 再帰的オートエンコーダ [Socher+ 11]
    - BiLSTMで文脈を考慮 [He+ 16]
  - 単語類似度行列のまとめ上げ
    - 各種プーリング
      [Socher+ 11][Yin+ 15][He+ 16]
  - 出力層: 回帰/Softmax

[He+ 16]



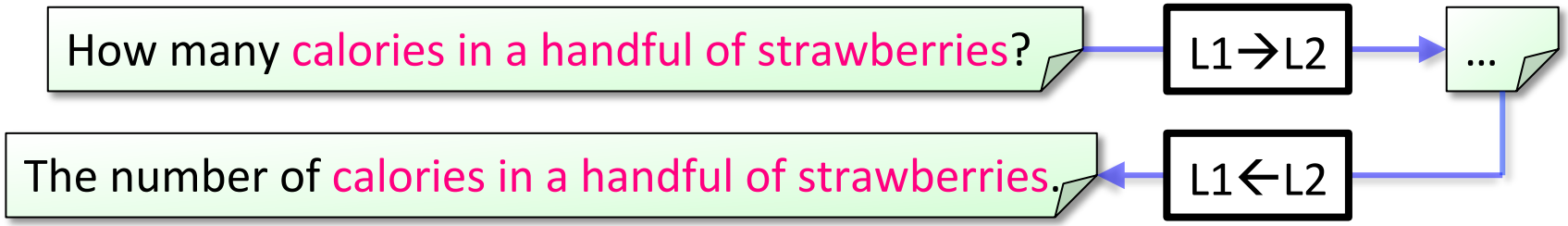
# 分散表現に基づくアプローチ (2/3)

- Siamese構造のCNNs [Yin+ 15]
  - 単語情報の抽象化
  - 多層化により構造的対応関係を表現
  - 動的プーリング+スタッキング
  - 事前学習
- Multi-perspective化 [He+, 15]
  - 類似度尺度
  - プーリング方法
  - CNNの形状



# 分散表現に基づくアプローチ (3/3)

- 折り返し翻訳としての尤度 [Mallinson+ 17]
  - $P(s_2|s_1) = \sum_t P(s_2|t)P(t|s_1)$
    - $P(t|s_1)$: L1→L2はn-bestデコード
    - $P(s_2|t)$: L2→L1は各tから$s_2$をforced decoding
  - NMTだと容易かつ滑らか



# 言い換えコーパス (文単位，正負例両方)

- 人間の手が入ったもの
  - 小説の複数訳 (英語) [Barzilay+ 01]
  - MT自動評価用の複数参照訳 (英語) [NIST 03][LDC 10;13]
  - MSRP: MSR ニュース言い換えコーパス (英語) [Dolan+ 04]
  - MSR ビデオ記述コーパス (英語，仏語) [Chen+ 11]
  - NTCIR RITE-2: 含意関係認識データ (日本語) [Watanabe+ 13]
  - WikiAnswers: 質問文 (英語) [Fader+ 13]
  - MS COCO: 画像キャプション (英語) [Lin+ 14]
  - SemEval STS: 意味的類似度コーパス (英語) [Agirre+ 14]
  - SemEval SICK: 構成的な文間関係データ (英語) [Marelli+ 14]
  - Twitter 言い換えコーパス (英語) [Xu+ 14][Lan+ 17]
  - Quora: 極めて大規模な質問文対 (英語) [Quora 17]

# 言い換えコーパス (文単位，ラベルなし)

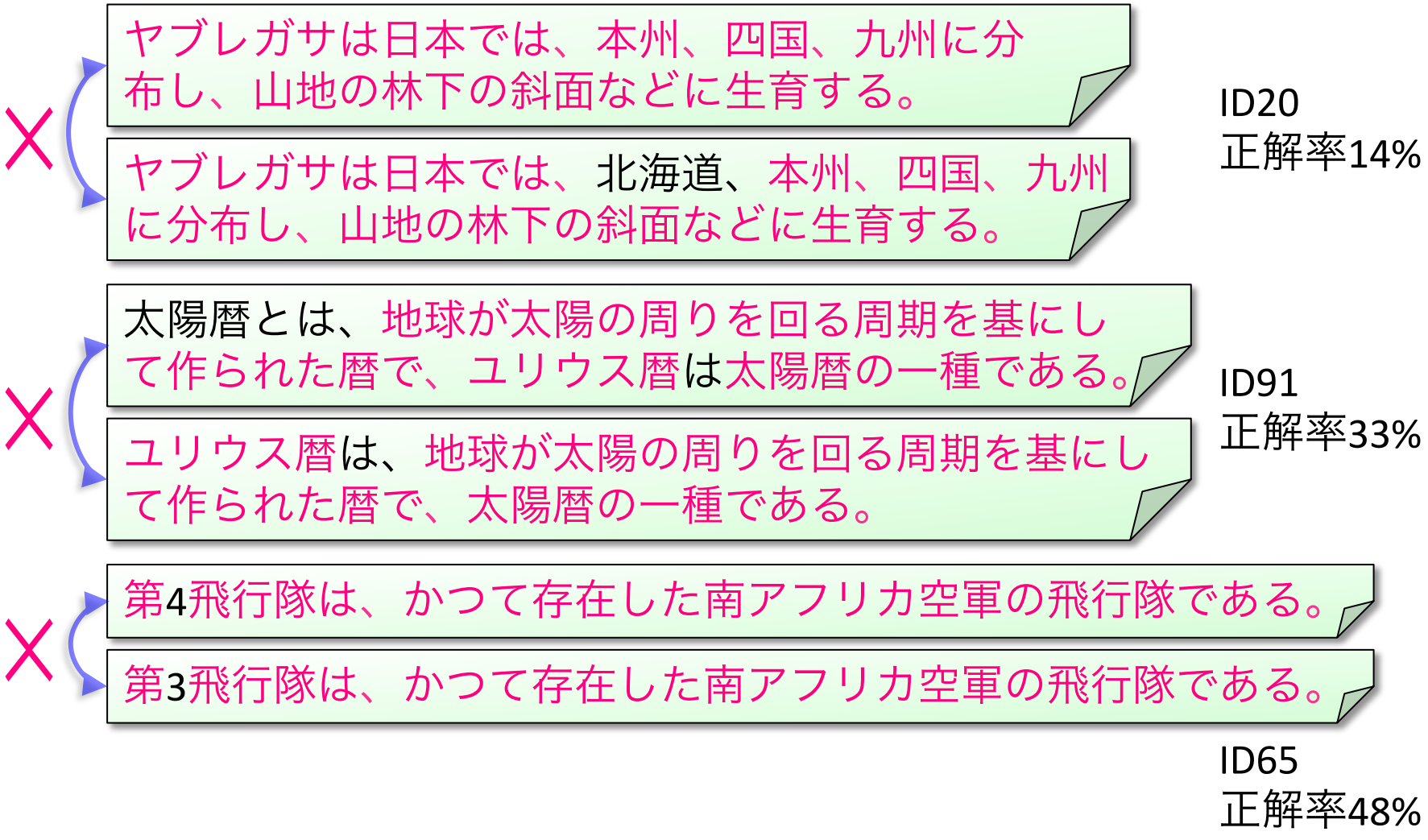
- 自動収集・生成されただけのもの
  - ParaNMT-50M: 人間訳とMT訳の対 [Wieting+ 17c]
  - OpusParCus: 多言語字幕からの抽出物 (6言語) [Creutz 18]

# これまでの研究の知見

- 表層的な手がかりが(十分に)有効
  - 表層的な類似度
  - 語彙的素性についてはスパースネス対策が有効
    - METEOR/TERpで局所的な言い換えを考慮 [Madnani+ 12]
    - Natural Logic & 隠れクラス [Watanabe+ 12][Guo+ 14]
    - 行列・テンソル分解 [Ji+ 13]
    - 分散表現 [Yin+ 15][He+ 15;16][Wieting+ 15;16;17a]
  - 統語構造やリッチな言語資源を用いても微々たる改善
- 実は誤謬
  - 標本バイアス: 実世界の問題の分布からの乖離 [Xu+ 14]
  - e.g., MSRP [Dolan+ 04]
    - 編集距離が8-20の文対
    - 評価用データの66.5%が正例 (常に「同義」でF値79.9)

# NTCIR RITE-2 [Watanabe+ 13]

- false negative の例



# NTCIR RITE-2 [Watanabe+ 13]

- false positive の例

筆箱とは、鉛筆、シャープペンシル、消しゴム、定規などを入れる物である。

筆箱は、鉛筆、消しゴム、定規、シャーペンなどを収めた箱だ。

ID292
正解率19%

サウスポーは、左利きのことである。

左利きの人を指す言葉として「サウスポー」がある。

ID86
正解率19%

忍者は、漫画のキャラクターとして頻繁に登場する。

忍者を題材とする漫画は、数多い。

ID26
正解率10%

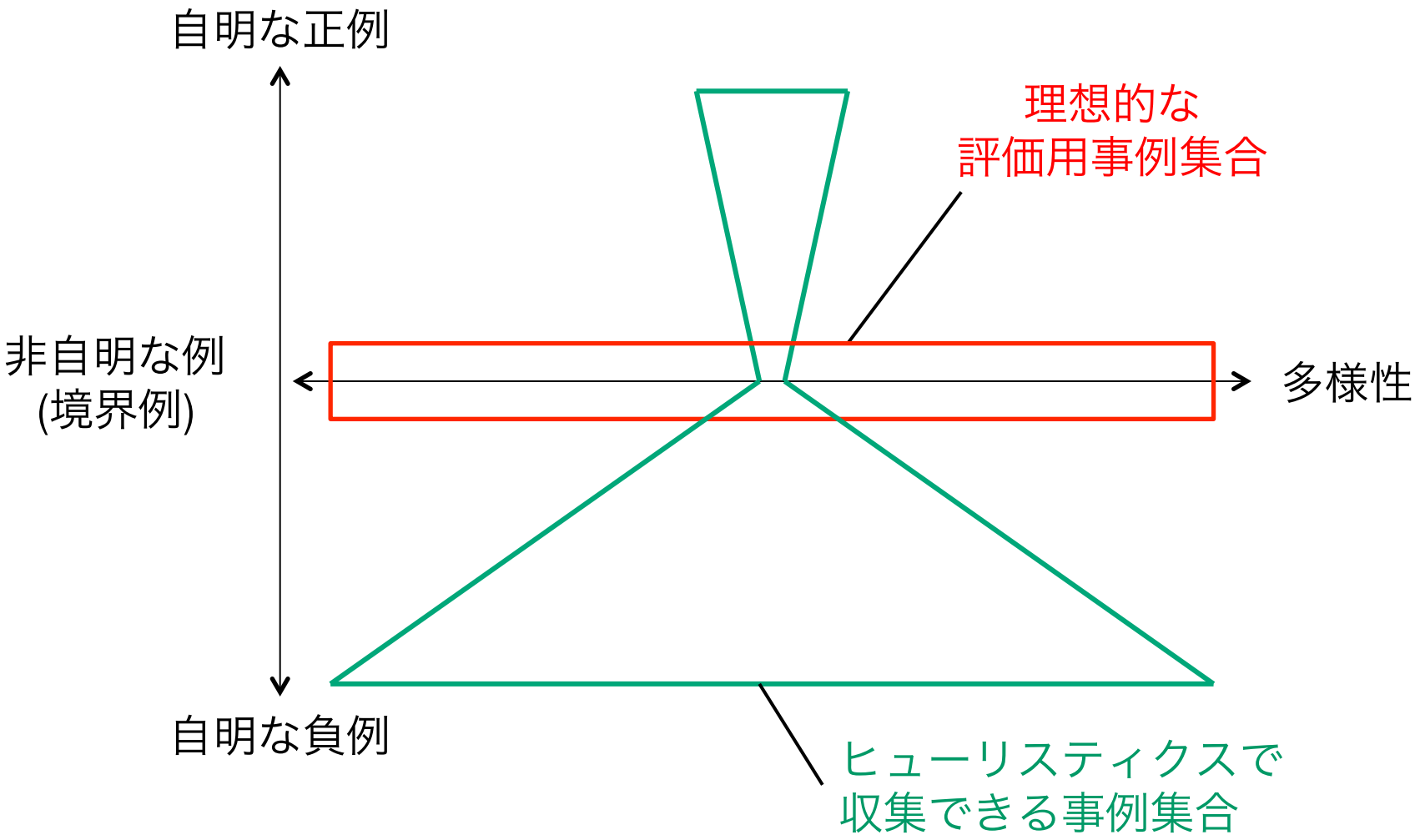
# 言い換え認識技術の評価用データの要件

- 3つの要件 [藤田+ 15]
  - 分布の自然さ: 目的から逆算して考える
    - 各応用でどんな言い換えを捉えたいか?
    - 分布も正負の基準も異なりうる [Dagan+ 05]
  - 正負例のバランス: 要議論
    - 自明な事例のみから学習できるか? (cf. 美術品の鑑定眼)
    - 境界をうまくとらえるような負例も必要 [Zaenen+ 05]
      - 言い換えの種類ごとの半自動事例生成/収集 [Fujita+ 05]
      - e.g., Winograd Schema Challengeの最小ペア [Levesque 11]
  - プリミティブさ: 担保可能 [Kaneko+ 13][藤田+ 15]
    - 独立な部分問題への分解と分類
      - 公平な評価: 惜しい誤答に部分点，まぐれ正解を減点
      - 一貫性のあるエラー分析が可能になる

# 評価用データの作り方の大まかな分類

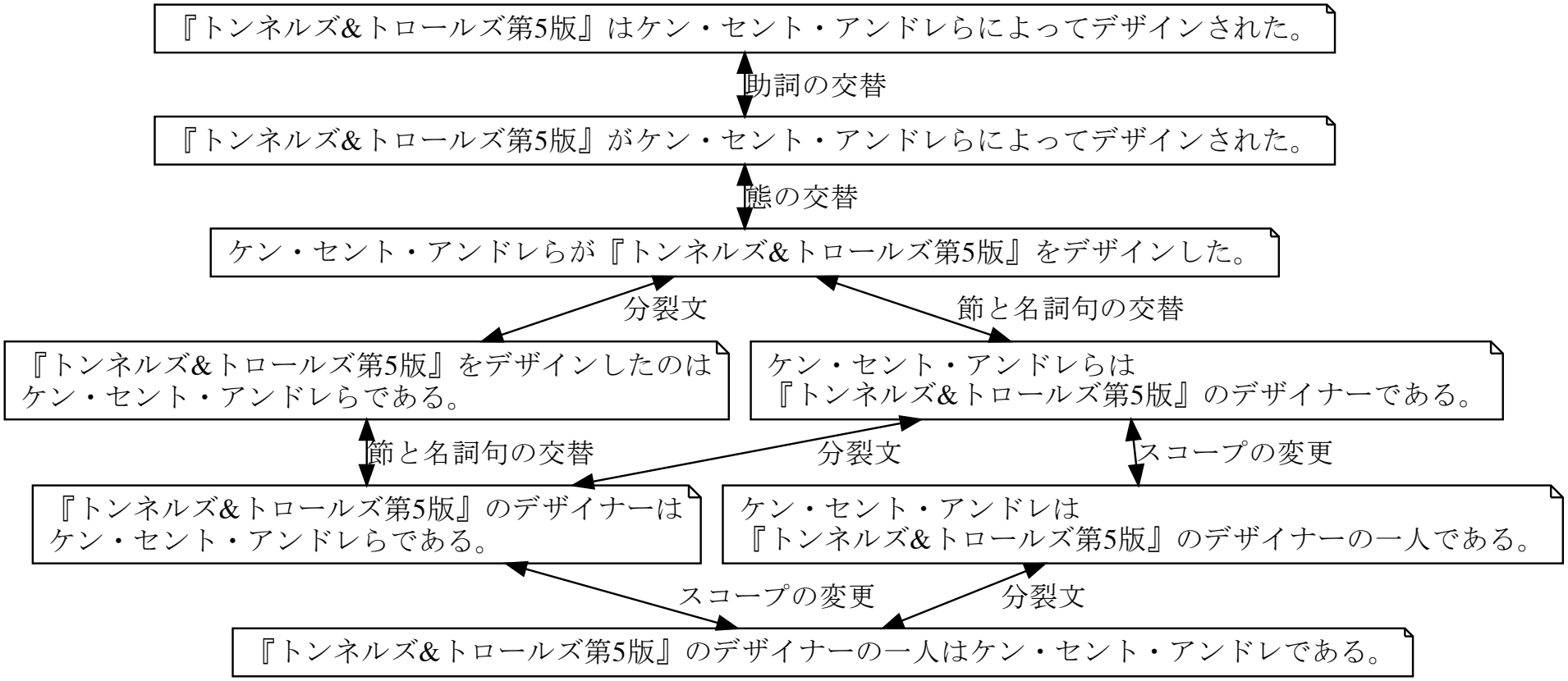
- ある範囲のテキストセットにアノテーション
  - 形態素解析，構文解析，固有表現抽出
  - 語義曖昧性解消，情報抽出
  - 述語項構造解析，照応解析，レビュー解析
  - 日本語校正，英語校正
- ある範囲の問題セットに模範解答を付与
  - MT，自動要約 (1つないし複数)
  - 情報検索 (見つけられる限り)
  - 言語生成，言い換え生成 (無数ではあるが)
- あたりをつけてから人手で選別 (言い換え認識)
  - 任意のテキスト対はほぼ間違いなく言い換えではない
  - 標本バイアスの回避 vs 人的コストの低減

# 分布の自然さと正負例のバランス



# NTCIR RITE-2の事例の分解例 [藤田+ 15]

- ID90-4



# データ作成時の精度にも問題

- 人間によるfalse positive

The stock rose $2.11, or about 11 percent, to close Friday at $21.51 on the New York Stock Exchange.

PG&E Corp. shares jumped $1.63 or 8 percent to $21.03 on the New York Stock Exchange on Friday.

Samsung halts production of its Galaxy Note 7 as battery problems linger.

#Samsung temporarily suspended production of its Galaxy #Note7 devices following reports

- 原因は何?
  - データ作成時のインストラクション?
  - 作業者の資質? [Xu+ 14][Lan+ 17]
  - RSTで作業者をスクリーニングできる?
- 言い換え生成の評価も同様

# もくじ

(5) pp.85-91 (7)

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(2) pp.34-52 (19)
(3) pp.54-65 (12)

(L4) 汎用的言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(4) pp.67-83 (17)

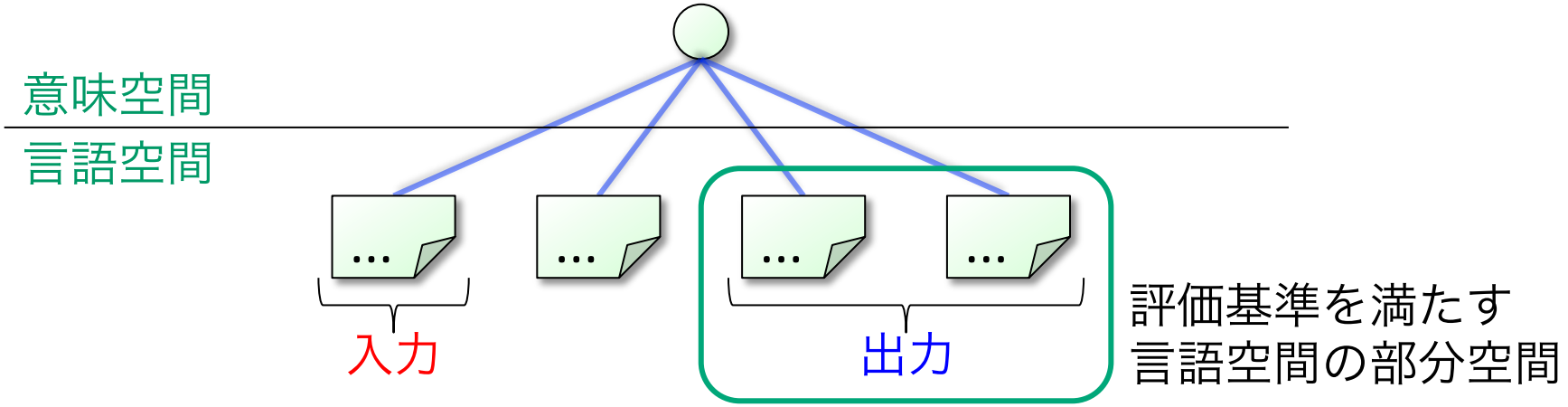
(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

(1) pp.17-32 (16)
(6) pp.93-112 (20)

# 言い換え生成

- タスクの定義
  - 入力: ある言語表現，目的に応じた評価基準
  - 出力: 入力と同じ意味，同じ言語の異なる言語表現の集合
    - ただし，与えられた評価基準を満たすもの




- 応用
  - 人間向け: 簡単化，文圧縮，音声合成(TTS)の前処理
  - 機械向け: MT前処理 (e.g., 未知語→既知語)

# 他のタスクとの関係

- text-to-text: MT, 文書要約, 簡単化
  - 共通点: 妥当な解は複数存在, 事前に網羅することは困難
  - MTとの違い: 言い換え生成はMTの特殊版
    - 単一の言語内の変換
    - 与えられた評価基準を満たす必要がある
    - 入力全体を書き換える必要はない
  - 文書要約, 簡単化との違い: 大部分が重複
    - 有用な言い換えとそうでないもの
    - 言い換え以外にも有用な操作, 処理がある
- 言い換え認識
  - 言い換え生成可能 → 言い換え認識可能
    - 逆は一般に不可能
    - cf. MT (生成) vs 対訳検索 (認識)

# 一般的な言い換え生成の手順

1. テキスト評価
   - 目的や制約に沿っていない箇所を同定
     - 簡単化: e.g., 語の難易度，構文パタンの難易度
     - MT前編集: e.g., 未知語の有無や数
     - 文圧縮: e.g., 文の長さ
2. 仮説生成
   - 必要な箇所を言い換えた候補を複数生成
     - 入力の一部を書き換える
     - 入力全体を書き換える: e.g., 単言語MT
3. テキスト評価
   - 目的や制約に照らして最適な候補を選択

# 過剰生成の例

- 「損害賠償を求める」 [Fujita+ 07]

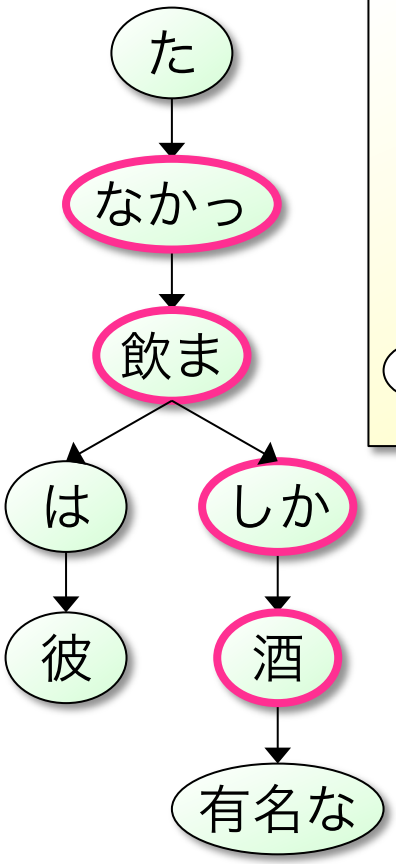
| | | | | |
|---|---|---|---|---|
| 損害が賠償される | 損害が賠償されたことが求める | 損害を賠償されたことが求める | 損害に賠償されたことが求める | 損害で賠償されたことが求める |
| 損害が賠償させる | 損害が賠償されたことを求める | 損害を賠償されたことを求める | 損害に賠償されたことを求める | 損害で賠償されたことを求める |
| 損害が賠償する | 損害が賠償されたことに求める | 損害を賠償されたことに求める | 損害に賠償されたことに求める | 損害で賠償されたことに求める |
| 損害を賠償される | 損害が賠償されたことで求める | 損害を賠償されたことで求める | 損害に賠償されたことで求める | 損害で賠償されたことで求める |
| 損害を賠償させる | 損害が賠償されることが求める | 損害を賠償されることが求める | 損害に賠償されることが求める | 損害で賠償されることが求める |
| 損害を賠償する | 損害が賠償されることを求める | 損害を賠償されることを求める | 損害に賠償されることを求める | 損害で賠償されることを求める |
| 損害に賠償される | 損害が賠償されることに求める | 損害を賠償されることに求める | 損害に賠償されることに求める | 損害で賠償されることに求める |
| 損害に賠償させる | 損害が賠償されることで求める | 損害を賠償されることで求める | 損害に賠償されることで求める | 損害で賠償されることで求める |
| 損害に賠償する | 損害が賠償させたことが求める | 損害を賠償させたことが求める | 損害に賠償させたことが求める | 損害で賠償させたことが求める |
| 損害で賠償される | 損害が賠償させたことを求める | 損害を賠償させたことを求める | 損害に賠償させたことを求める | 損害で賠償させたことを求める |
| 損害で賠償させる | 損害が賠償させたことに求める | 損害を賠償させたことに求める | 損害に賠償させたことに求める | 損害で賠償させたことに求める |
| 損害で賠償する | 損害が賠償させたことで求める | 損害を賠償させたことで求める | 損害に賠償させたことで求める | 損害で賠償させたことで求める |
| 賠償が損害される | 損害が賠償させることが求める | 損害を賠償させることが求める | 損害に賠償させることが求める | 損害で賠償させることが求める |
| 賠償が損害させる | 損害が賠償させることを求める | 損害を賠償させることを求める | 損害に賠償させることを求める | 損害で賠償させることを求める |
| 賠償が損害する | 損害が賠償させることに求める | 損害を賠償させることに求める | 損害に賠償させることに求める | 損害で賠償させることに求める |
| 賠償を損害される | 損害が賠償させることで求める | 損害を賠償させることで求める | 損害に賠償させることで求める | 損害で賠償させることで求める |
| 賠償を損害させる | 損害が賠償したことが求める | 損害を賠償したことが求める | 損害に賠償したことが求める | 損害で賠償したことが求める |
| 賠償を損害する | 損害が賠償したことを求める | 損害を賠償したことを求める | 損害に賠償したことを求める | 損害で賠償したことを求める |
| 賠償に損害される | 損害が賠償したことに求める | 損害を賠償したことに求める | 損害に賠償したことに求める | 損害で賠償したことに求める |
| 賠償に損害させる | 損害が賠償したことで求める | 損害を賠償したことで求める | 損害に賠償したことで求める | 損害で賠償したことで求める |
| 賠償に損害する | 損害が賠償することが求める | 損害を賠償することが求める | 損害に賠償することが求める | 損害で賠償することが求める |
| 賠償で損害される | 損害が賠償することを求める | 損害を賠償することを求める | 損害に賠償することを求める | 損害で賠償することを求める |
| 賠償で損害させる | 損害が賠償することに求める | 損害を賠償することに求める | 損害に賠償することに求める | 損害で賠償することに求める |
| 賠償で損害する | 損害が賠償することで求める | 損害を賠償することで求める | 損害に賠償することで求める | 損害で賠償することで求める |

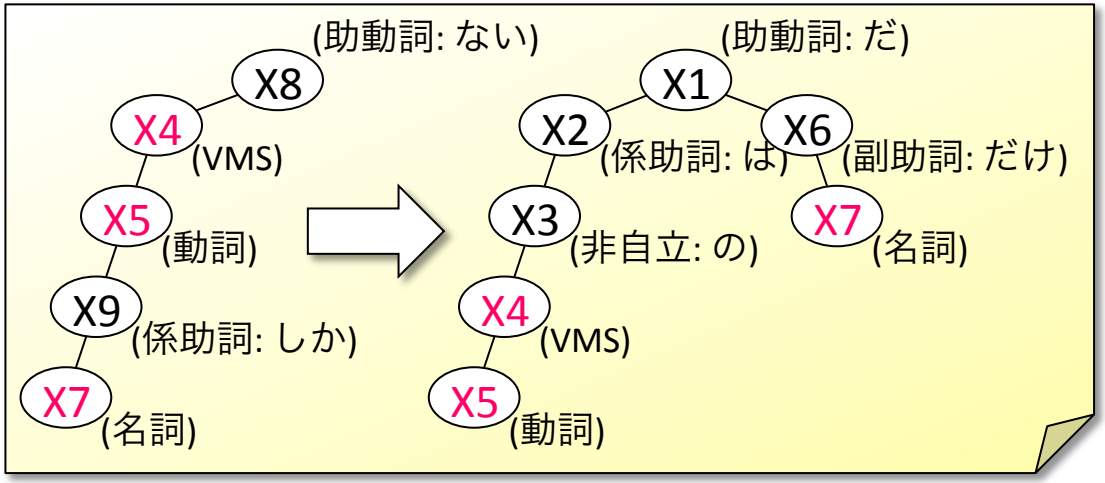
| | | | | |
|---|---|---|---|---|
| 損害賠償が求める | 損害が求める | 損害された賠償が求める | 損害させた賠償が求める | 損害した賠償が求める |
| 損害賠償に求める | 損害を求める | 損害された賠償を求める | 損害させた賠償を求める | 損害した賠償を求める |
| 損害賠償で求める | 損害に求める | 損害された賠償に求める | 損害させた賠償に求める | 損害した賠償に求める |
| | 損害で求める | 損害された賠償で求める | 損害させた賠償で求める | 損害した賠償で求める |
| 損害の賠償が求める | 賠償が求める | 損害される賠償が求める | 損害させる賠償が求める | 損害する賠償が求める |
| 損害の賠償を求める | 賠償を求める | 損害される賠償を求める | 損害させる賠償を求める | 損害する賠償を求める |
| 損害の賠償に求める | 賠償に求める | 損害される賠償に求める | 損害させる賠償に求める | 損害する賠償に求める |
| 損害の賠償で求める | 賠償で求める | 損害される賠償で求める | 損害させる賠償で求める | 損害する賠償で求める |

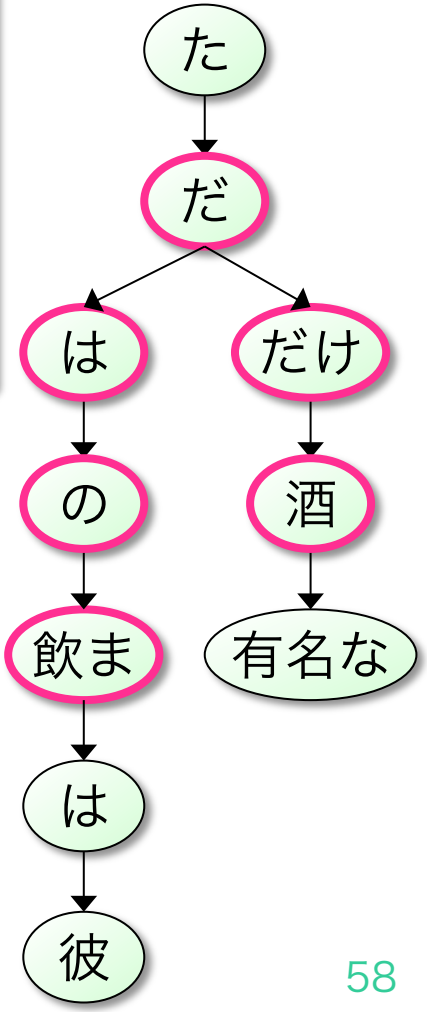
青字のみ文法的な表現，赤枠のみ同義の表現

# 仮説生成手法の例 (1/4)

- 統語構造の変換 [Lavoie+ 00][Takahashi+ 01][Zhu+ 10]
  - 必要になる後処理の分類・体系化 [藤田+ 03]

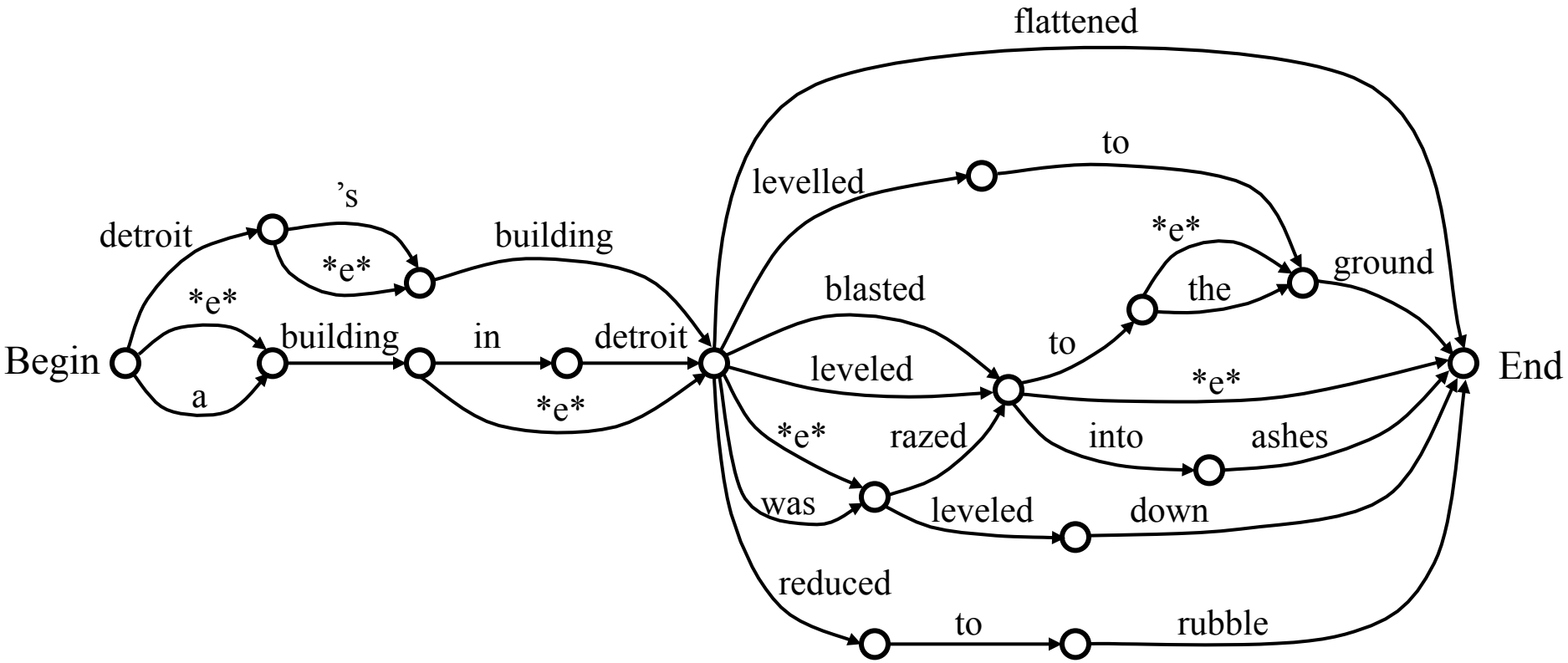


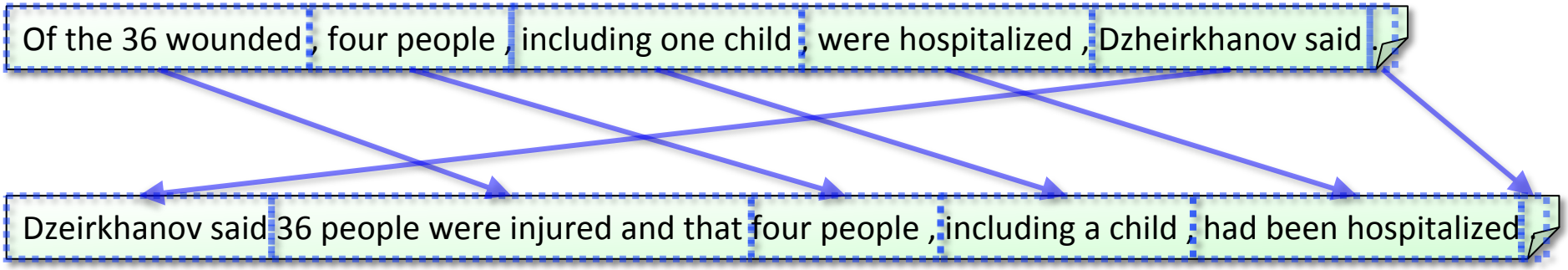
# 仮説生成手法の例 (2/4)

- 単言語パラレルデータ → オートマトン
  - Multiple-sequence alignment [Barzilay+ 03]

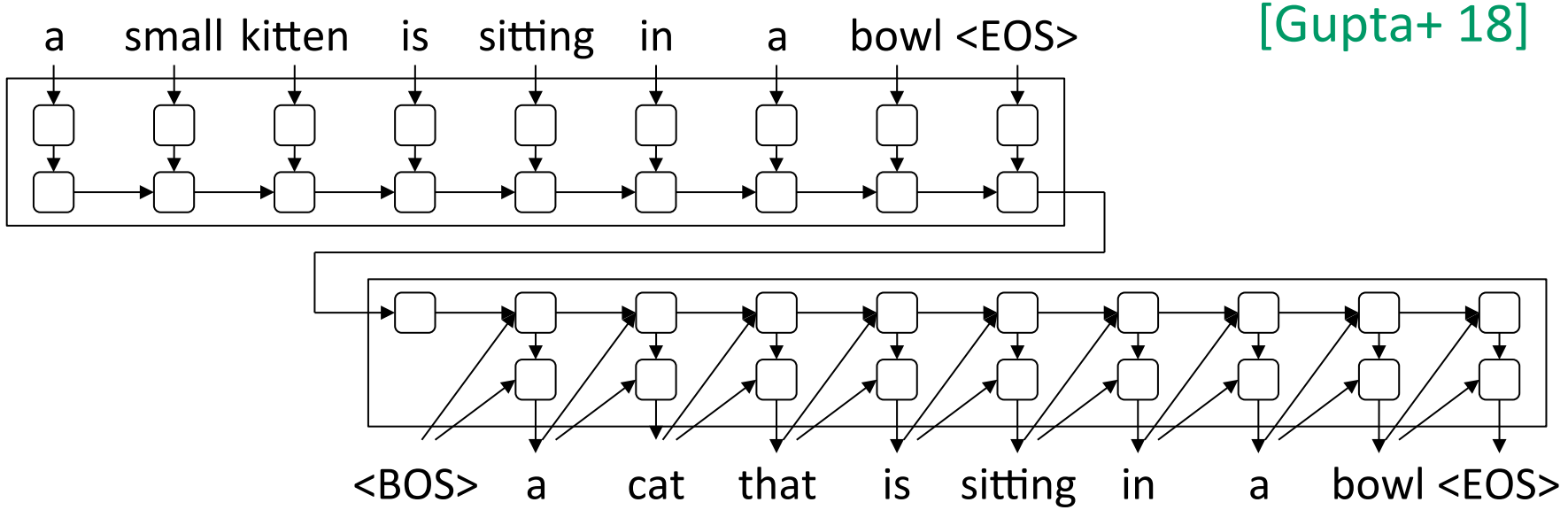


# 仮説生成手法の例 (3/4)

- 単言語パラレルデータに対する単言語MT
  - 句に基づく統計的MT (PBSMT) [Quirk+ 04][Wubben+ 10]

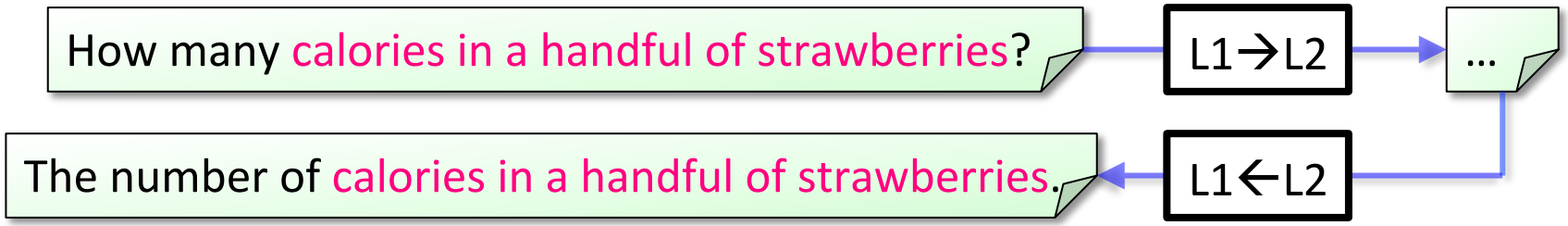



  - ニューラルMT (NMT) [Prakash+ 16][Cao+ 17][Li+ 17] [Gupta+ 18]



# 仮説生成手法の例 (4/4)

- MTによる折り返し翻訳 [Sun+ 12][Mallinson+ 17]
  - 単言語パラレルデータは小規模にしか存在しない
    → 対訳コーパスを活用
  - 人間の評価との相関: SMT < NMT



# 言い換えの仮説生成手法の比較

- 入力の一部を書き換える手法
  - ✓ 事例研究の成果を取り込める可能性
  - ✗ 複数の言い換えの組み合わせ制御が困難
- 入力全体を書き換える手法
  - ✓ 既存のツールを援用可能 (e.g., MTのツール)
  - ✓ フレームワークの数学的背景
  - ✗ 特定の一部分のみを書き換えるという制御が困難

# テキスト評価

- 評価基準(目的や制約)に照らして最適な候補を選択
  - 簡単化: e.g., 語の難易度, 構文パタンの難易度
  - MT前編集: e.g., 未知語の有無や数
  - 文圧縮: e.g., 文の長さ
- 書き換えた箇所とその他の部分の整合性の検証
  - 変換知識の適用条件は一般的に不完全 (過剰生成が前提)
  - 入力全体を生成し直した場合も
- アプローチ
  - フィルタリング: 禁則処理など
  - ランキング: 候補間の比較 (cf. MTのn-bestリランキング)
  - 自動訂正: 活用・屈折の修正や助詞の修正など

# 言い換え生成の評価

- 出力を得るたびに人間が評価
  - 生じうる誤りの全体像 [藤田+ 03]
  - MTの評価基準に倣ったライカート尺度 [Callison-Burch 08]
    - 文法性 (流暢さ): 5段階
    - 原文との意味の等価性: 5段階
  - 数値評価ではなく決定木による分類 [Fujita 13c]
- 少数の参照出力との照合による自動評価: cf. MT
  - MT向け自動評価尺度の援用: BLEU, METEOR, TER
  - 言い換え向けに提案されたもの
    - PEM: MT結果を「意味」とみなして照合 [Liu 10]
    - PINC: 参照出力ではなく原文との非類似度 [Chen+ 11]
    - iBLEU: BLEUと(1-PINC)の線形補間 [Sun+ 12]

# 言い換え生成の課題

- 効率的かつ再現性のある評価手法の確立
  - 出力を得るたびに人間による評価が必要
    - 妥当な解は複数存在，事前に網羅することは困難
    - 自動評価は妥協案?
  - 結局は目的依存
    - 文章執筆支援: 候補を提示できればあとは人間任せ
    - MT前編集: 非文法的でもなんとかなる
- 仮説生成部の改良
  - 頑健性: 多様な言い換えの実現
    - 言い換えの種類ごとに適した手法の高度化と統合
    - 必要な語彙知識の整備
    - ニューラル言い換え生成の未来
  - 正確性: テキスト評価に頼り過ぎない
    - 語彙意味論などの深い知識の導入

# もくじ

(5) pp.85-91 (7)

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(2) pp.34-52 (19)　(3) pp.54-65 (12)

(L4) 汎用的言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(4) pp.67-83 (17)

(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

(1) pp.17-32 (16)　(6) pp.93-112 (20)

# 言い換えの処理に必要な知識 [藤田+ 04]

- 必要性が論じられているもの
  - 語の素性/意味記述
    - e.g., 項構造/意味役割，動詞の使役・受身の可否
  - 表現間の関係に関する知識
    - e.g., 同義の語句の辞書 (言い換え知識)
      - 言い換え知識の自動獲得 (paraphrase acquisition)

語 窮地 ↔ 苦境　重傷 ↔ 大ケガ

句 脚光を浴びる ↔ 注目される　重傷を負う ↔ 大ケガをする

    - e.g., 派生語，反義語，特質構造，上位-下位，部分-全体
- 全貌はまだまだ不明
  - 言い換えの種類(L1/L2)と一緒に考える必要あり [藤田+ 04]
  - End-to-Endのアプローチでは今のところ無視されがち
    - 本当に不要かはまだ分からない

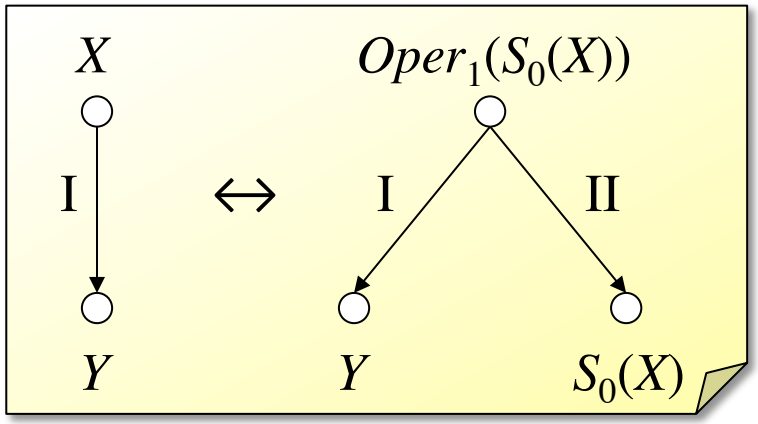
# 古典的な言い換え知識

- 人手で記述された構文的な言い換え規則
  - 変形文法 [Harris 81]

NP1 V1 (+AUX) V2 (-AUX) NP2
→ NP2 V1 BE V2-PP by NP1

能動態 → 受動態

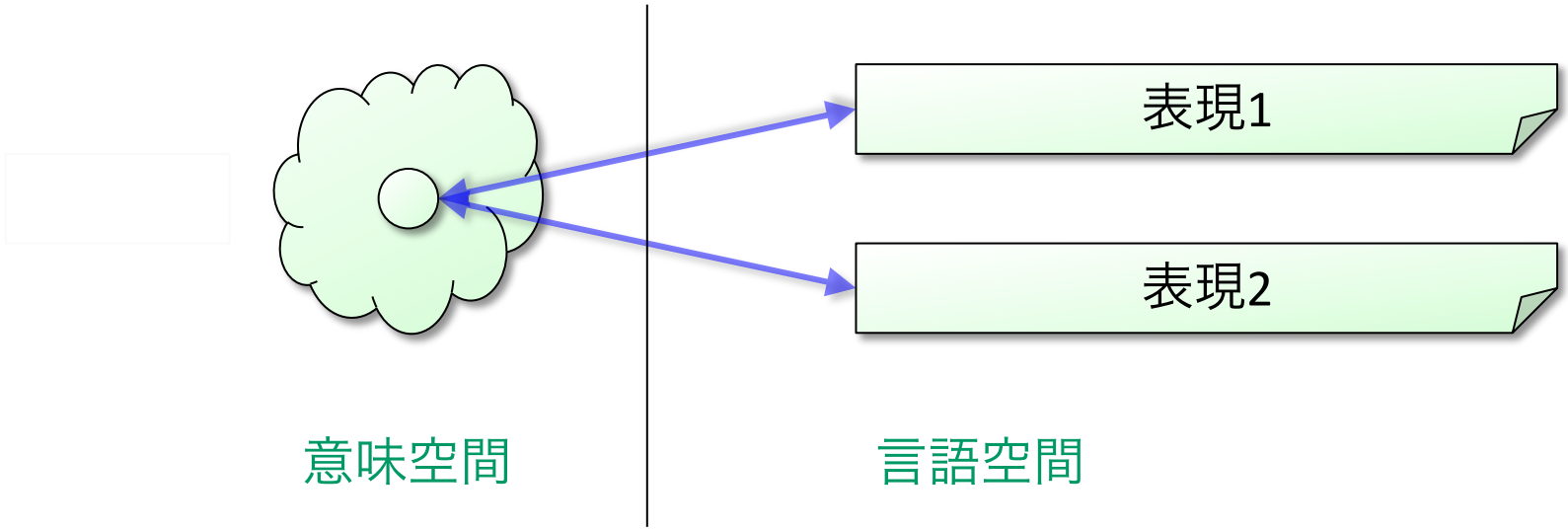
  - Meaning-Text Theory [Mel'čuk+ 87]




動詞句 ↔
機能動詞構文

# 言い換え知識

- 何らかの手がかりで同義性を担保する

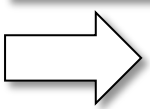



- 人間が設定した概念ラベル: Synset ID, 見出し語
- 文脈: bag-of-(周辺の表現), bag-of-(対訳フレーズ)
- 同一の情報源: 原文書, イベント, 動画
- etc.

# 人間が設定した概念ラベル (1/2)

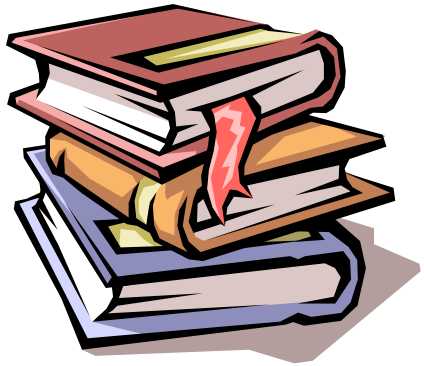
- 同じ概念IDを持つ語句
  - e.g., WordNet [Miller+ 85]

02526085: achieve, accomplish, attain, reach
05793554: basis, base, cornerstone, foundation, ...



achieve ↔ accomplish

base ↔ basis



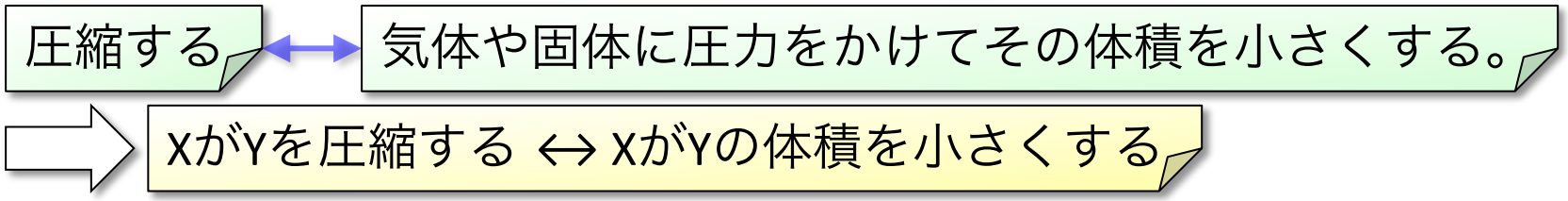
  - 最大の課題: 人手が必要
    - e.g., 新語, 固有表現

google (v) ↔ search the Web using Google

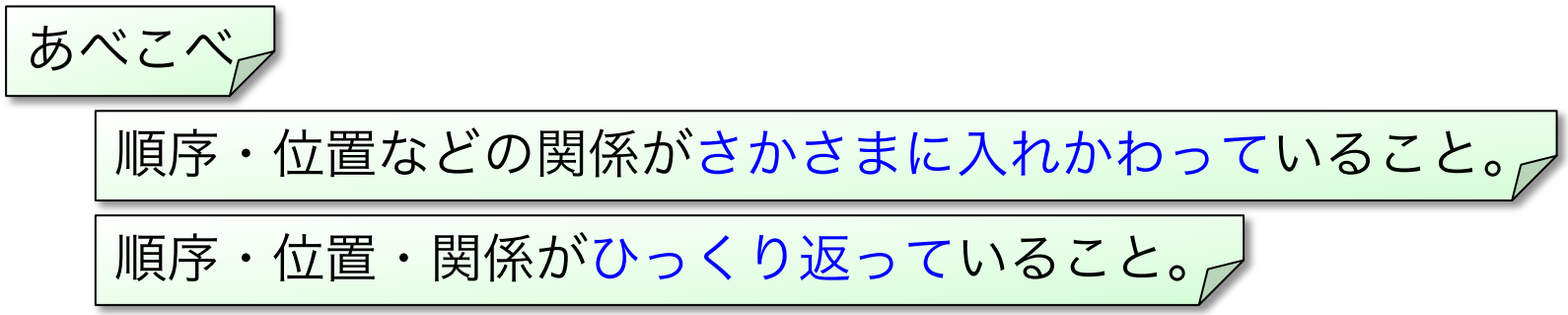
先進的音声翻訳研究開発推進センター ↔ ASTREC

# 人間が設定した概念ラベル (2/2)

- 見出し語と語釈文の対応付け
  - 各種整形 [藤田+ 00][梶原+ 13]
  - 項構造の対応付け [Kaji+ 02][青山+ 07][Sasano+ 13]




- 同じ語に対する複数の辞書の語釈文 [村田+ 04]




- 課題: スケーラビリティと恣意性
  - 人間向けの辞書の存在が前提
  - Webから定義文を拾ってくる [Hashimoto+ 11]

# 言い換え知識の自動獲得

- タスクの定義
  - 入力: コーパス等の言語資源
  - 出力: 同義表現の集合(一般に対)の集合
    - ただし文よりも小さなテキスト断片とすることが多い “sub-sentential paraphrases”
- 知識の規模
  - 変形ルールや表現対の手書き ($10^2$〜$10^4$)
  - コーパスからの自動獲得 ($10^6$〜$10^9$)

# 文脈: Bag-of-(周辺の表現)

- 分布仮説 [Harris 54][Firth 57]
  - “類似する文脈でよく使われる表現は似た意味を持つ”
  - e.g., 単言語コーパス中の左右のn-gram, 構造上の隣接要素
  - e.g., “Tezgüno” [Pantel+ 02]
    - wine, cognac, whiskey と似ている → アルコール飲料

A bottle of tezgüno is on the table

Everyone likes tezgüno

Tezgüno makes you drunk

We make tezgüno out of corn

  - 必ずしも同義とは限らない

同義 resembles ↔ looks like

反義 decrease ↔ increase

類義 tezgüno ↔ wine

上位・下位 liquor ↔ wine

# 文脈: Bag-of-(対訳フレーズ)

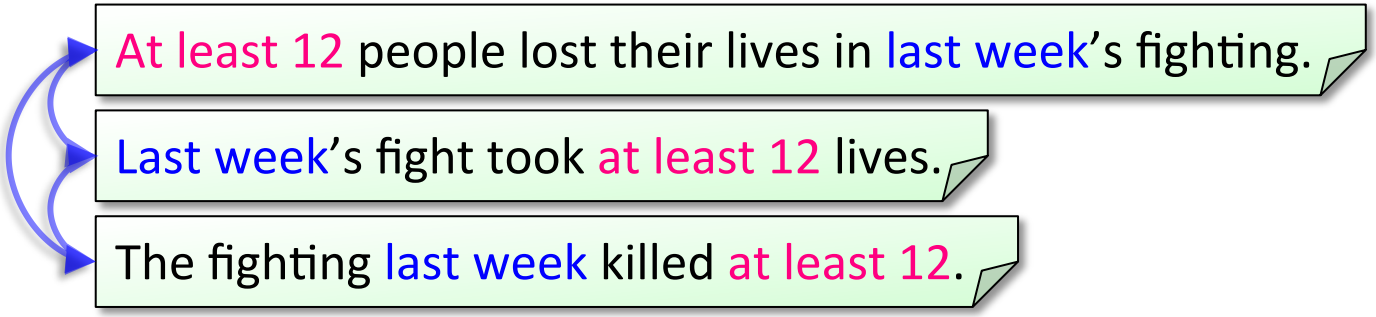
- 他の言語の表現を意味表現とみなす [Bannard+ 05]
  - “pivoting”
  - 分布仮説における「文脈」よりも信頼できる
  - 対訳コーパスから自動的に獲得できる
    - i.e., 単語アラインメント + 対訳フレーズ対の抽出
      $P(s_2|s_1) = \sum_t P(s_2|t)P(t|s_1)$

自動的に学習した翻訳テーブル

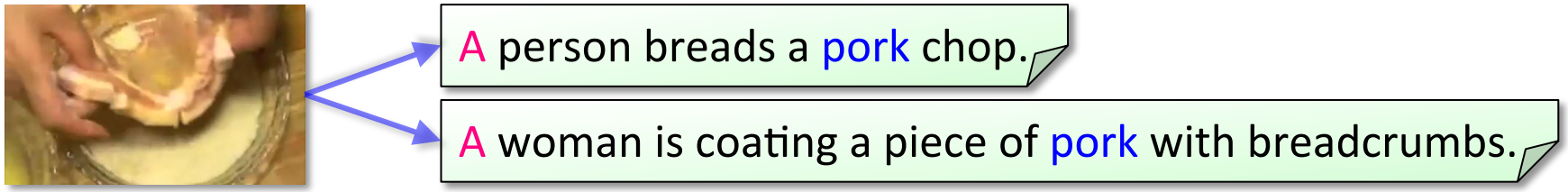
| | | |
|---|---|---|
| thrown into jail | ||| | festgenommen |
| imprisoned | ||| | festgenommen |
| under control | ||| | unter kontrolle |
| in check | ||| | unter kontrolle |

# 同一の情報源に対する異なる言語化

- 単言語パラレルコーパス
  - 数式の言語化 [Barzilay+ 02]
  - 同一の原文書に対する複数の翻訳
    [Barzilay+ 01][Pang+ 03][Wieting+ 17b][Arase+ 17][Creutz 18]




- 単言語コンパラブルコーパス
  - 画像や動画の内容の言語化 [Chen+ 11][Lin+ 14]

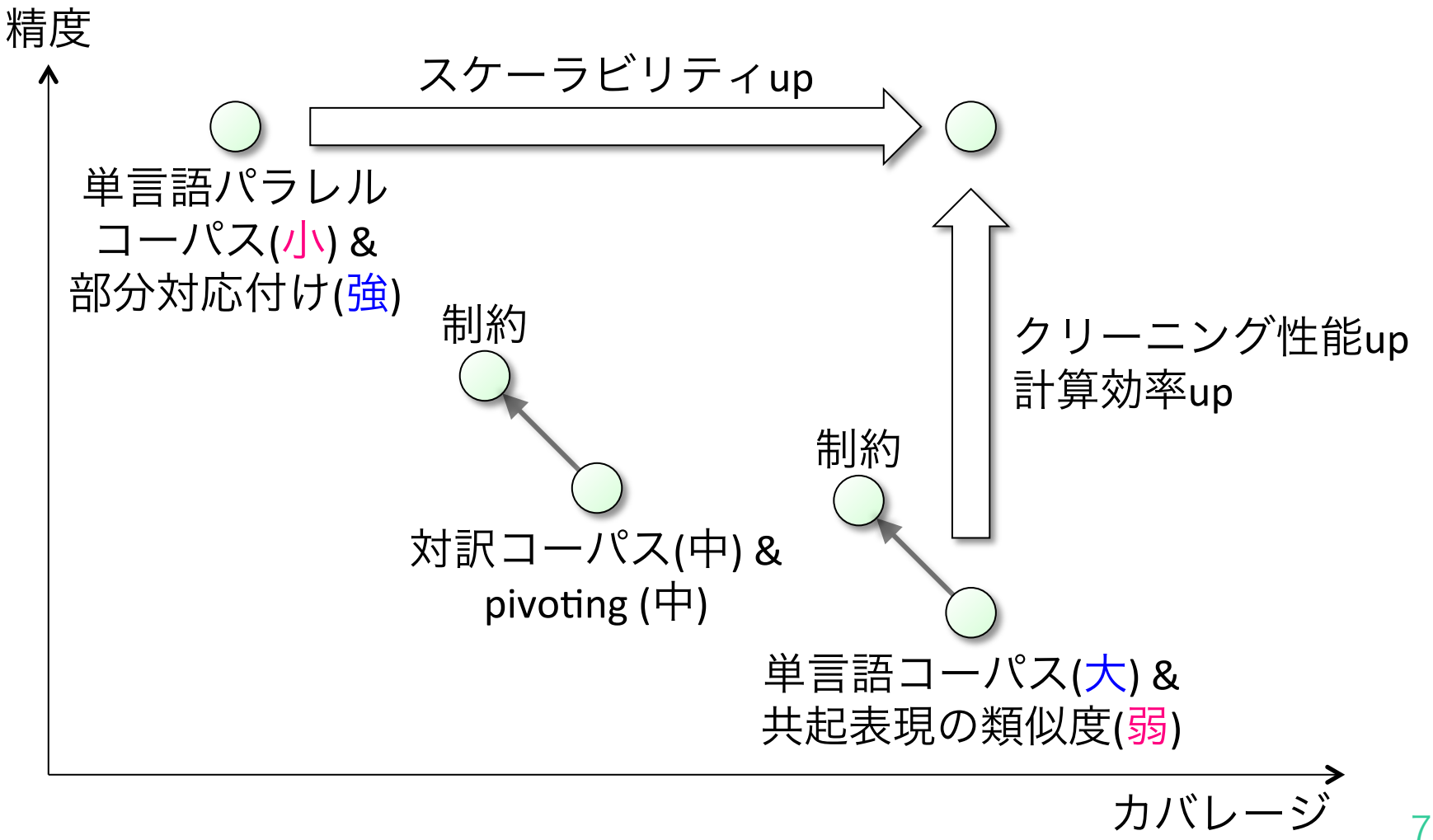


# 同一のイベントについて述べた複数の記事

- 単言語コンパラブルコーパスの自動構築
  - 同じ内容とは限らないが類似する情報を含む
  - 異なる通信社の記事
    - 編集距離に基づくクリーニング [Dolan+ 04]
  - Google News の見出しの対応付け [Wubben+ 09]
  - 同じ語に対する複数の定義文の獲得 [Hashimoto+ 11]
  - 同じトピックやURLに言及するtweets [Xu+ 14][Lan+ 17]
  - QAサイトの関連する質問文 [Fader+ 13][Quora 17]
- 対応箇所の抽出
  - 細粒度の固有表現体系 + 分布仮説 [Shinyama+ 02]
  - Multiple-sequence alignment [Barzilay+ 03]
  - 統語構造上の部分木の対応付けと評価 [Hashimoto+ 11][Arase+ 17]

# 言い換え知識獲得における課題

- カバレージと精度の両立
  - 制約の導入 = 上澄み抽出



# スケールしそうなアプローチ (1/2)

- Web → 単言語コンパラブルコーパス → 知識抽出
  - Webから得られる複数の定義文の部分対応付け [Hashimoto+ 11][Yan+ 13]
    1. 定義文らしさの判定
    2. 依存構造木上の全部分木の列挙
    3. 対応する動詞句の発見 (対応するか否かの判定)

骨粗鬆症

骨粗鬆症とは、骨の量が減り、骨がもろくなってしまう病気だ。

骨粗鬆症とは、骨量が低下し、骨が折れやすくなる病気だ。

骨の量が減る ↔ 骨量が低下する

骨がもろくなってしまう ↔ 骨が折れやすくなる

# スケールしそうなアプローチ (2/2)

- 高品質な知識を単言語コーパスで拡張 [Fujita+ 15]

airports in Europe ↔ Europeanan airports
amendment of regulation ↔ amending regulation
should be noted that ↔ is worth noting that

1. 知識の抽象化 (事例のパターン化)

*X*:ϕ in *Y*:ϕ ↔ *Y*:an *X*:ϕ
*X*:ment of *Y*:ϕ ↔ *X*:ing *Y*:ϕ
should be *X*:ed that ↔ is worth *X*:ing that

2. 具体例の抽出

cohesion in Europe ↔ Europeanan cohesion
democracy in Europe ↔ Europeanan democracy
increase in Haiti ↔ Haitianan increase
transportation in suburb ↔ suburbanan transportation
economy in Uruguay ↔ Uruguayanan economy
amendment of documents ↔ amending documents
amendment of protocol ↔ amending protocol
investment of resources ↔ investing resources
recruitment of engineers ↔ recruiting engineers
should be highlighted that ↔ is worth highlighting that
should be reiterated that ↔ is worth reiterating that
should be stated that ↔ is worth stating that

# 言い換え知識の自動獲得 (1/2)

- [Barzilay+ 01] 自動獲得の草分け
  - 単言語パラレルコーパス (小説の複数訳，4.5万文対)
    → 英語，9,483表現対，文脈なしの精度92%
- [Bannard+ 05] “pivoting”初登場 → 標準法として定着
  - 多言語対訳コーパス (Europarl 英-独西，103万文対)
    → 英語，数については言及なし，文脈なしの精度65%
    → 文脈ありの精度30% [Callison-burch 08]
- [Callison-burch 08] CCG制約の導入
  - 多言語対訳コーパス (Europarl 英-10言語，160万文対)
    → 英語，370万表現対(*)，文脈ありの精度55%
- [Ganitkevitch+ 13] コーパスの拡大，文脈類似度の付加
  - 多言語対訳コーパス (英語-各種欧州言語，1.06億文対)
    → 英語，7,600万表現対(*) + 9,360万パターン対
  - 多言語対訳コーパス (スペイン語-各種欧州言語，1,500万文対)
    → スペイン語，1.06億表現対(*) + 5,530万パターン対

(*) 適用条件ごとに1件

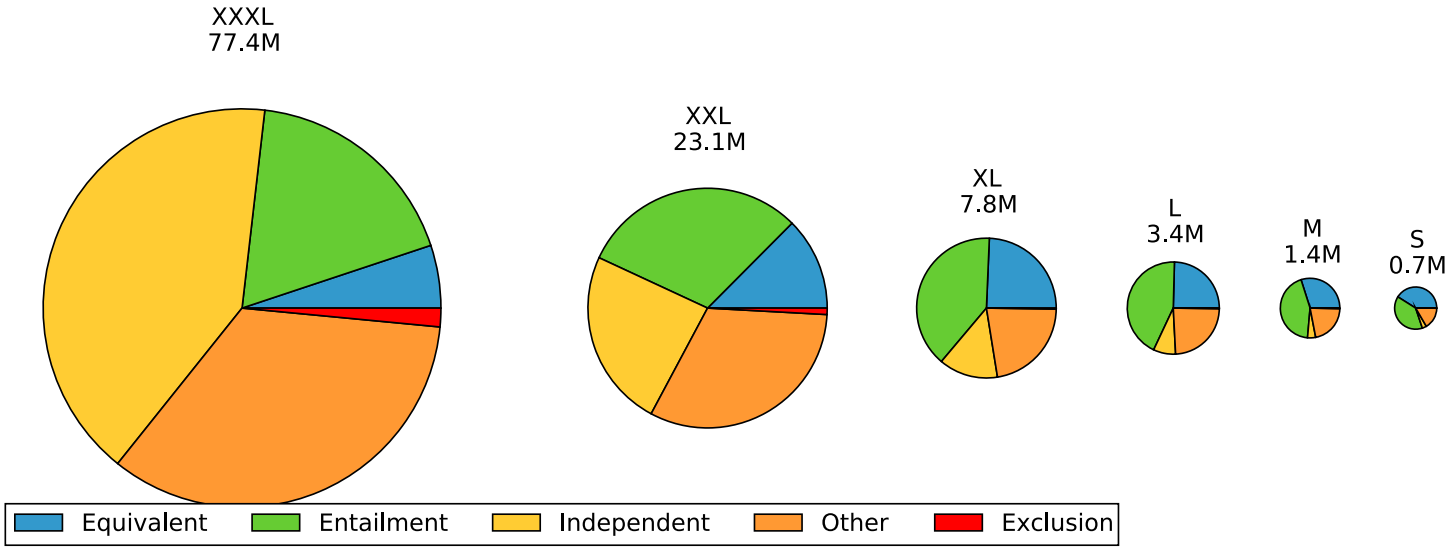
create equal: VP/(NP/NNS) ↔ creating equal

# 言い換え知識の自動獲得 (2/2)

- [De Saeger+ 09] 大規模コーパス&名詞クラス化
  - 単言語コーパス (Web文書，5000万文書)
    → 日本語，170万パターン (に対して1つ以上の言い換え)
- [Hashimoto+ 11] 大規模コンパラブルコーパスの自動獲得
  - 単言語コーパス (Web文書，6億文書)
    → 日本語コンパラブルコーパス3,000万定義文対
    → 日本語，30万表現対 (推定)，文脈ありの精度94%
- [Mizukami+ 14] “pivoting”を日本語に適用
  - 2言語対訳コーパス (日英, 2400万語)
    → 日本語，4,700万表現対
- [Fujita+ 15] 大規模コーパスを用いた語彙的拡張
  - 2言語対訳コーパス (英仏，200万文対)
    + 単言語コーパス (報道記事，5,200万文)
    → 英語，6,380万表現対 + 25.2万パターン対，文脈ありの精度61%

# 言い換え知識獲得のもう一つの課題

- 精密化
  - 言い換えが成立する条件
    - e.g., CCG制約 [Callison-burch 08]
    - e.g., 左右n-gram [Marton 09]
  - “phrase embedding”
  - 表現対の意味的関係 [Pavlick+ 15]
    - 置換可能な文脈があるものは含めておきたいが…



# 言い換え知識ベース等

- 人間の手が入った言い換え知識等
  - WordNet [Miller+ 93]
  - 日本語WordNet [Bond+ 12]
  - 日本語機能表現辞書「つつじ」 [松吉+ 08]
  - ALAGIN動詞含意関係DB [Hashimoto+ 09;11]
  - ALAGIN日本語異表記対DB [Kojima+ 10]
- 自動獲得された言い換え知識
  - ALAGIN日本語パターン言い換えDB [De Saeger+ 09]
  - TERp 付属のデータ (英語) [Snover+ 09]
  - METEOR 付属のデータ (6言語) [Denkowski+ 11]
  - PPDB: CCG制約つき言い換えDB (21言語) [Ganitkevitch+ 13]
  - PPDB-Ja: 日本語言い換えDB [Mizukami+ 14]
  - Lexpanded PPDB: 語彙的に拡張したPPDB (4言語) [Fujita 16]

# もくじ

(5) pp.85-91 (7)

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(2) pp.34-52 (19)
(3) pp.54-65 (12)

(L4) 汎用的言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(4) pp.67-83 (17)

(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

(1) pp.17-32 (16)
(6) pp.93-112 (20)

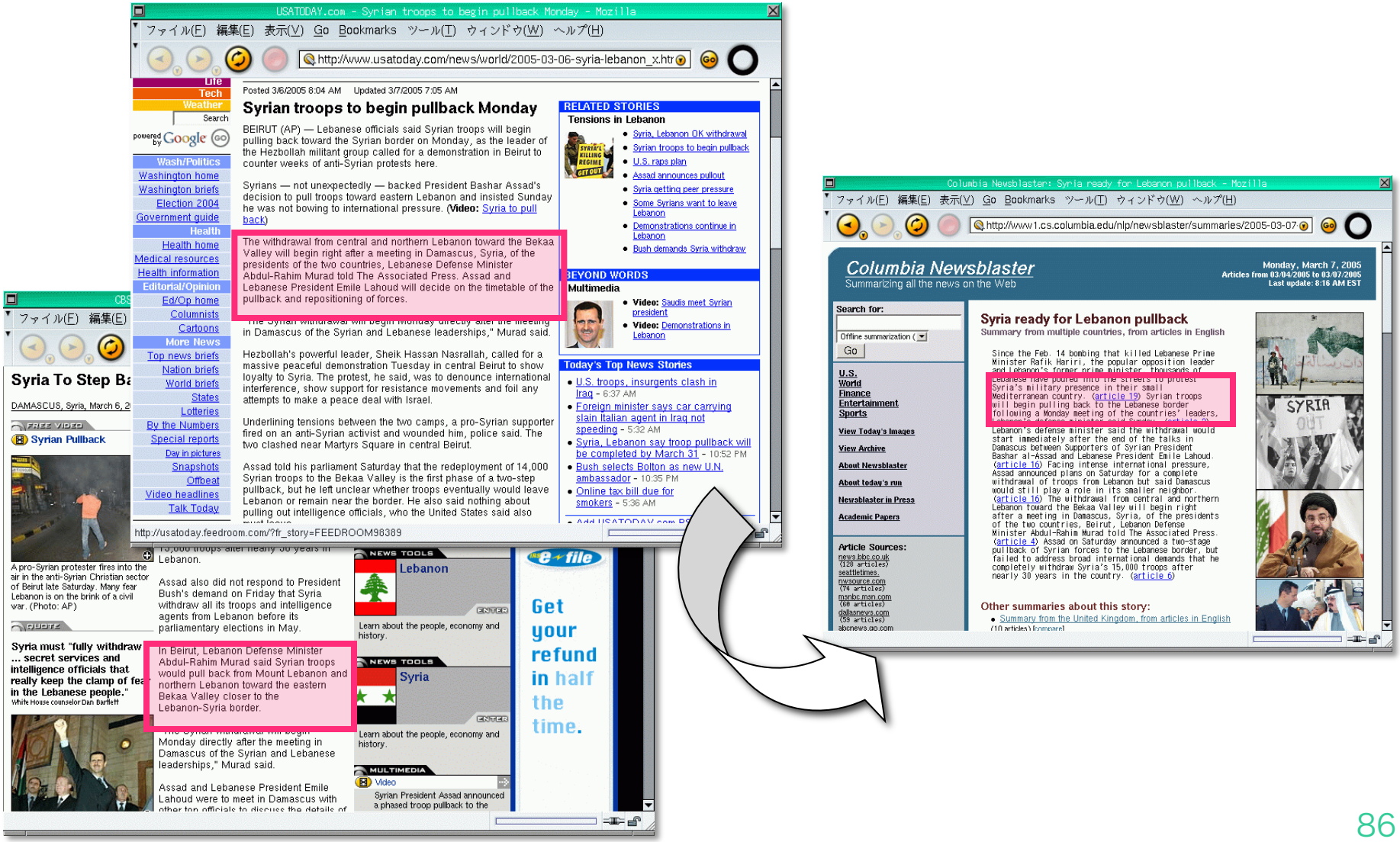
# 応用技術と言い換え

| | 言い換え生成 | 言い換え認識 |
|---|---|---|
| 人間向け | 文章執筆支援<br>文章読解支援<br>TTS前処理<br>MT後編集<br>ステガノグラフィ | 剽窃検出<br>翻訳メモリ検索<br>情報検索<br>質問応答<br>東口ボ |
| 機械処理向け | SMTのモデル拡張<br>MT前編集 | データマイニング |

文書要約

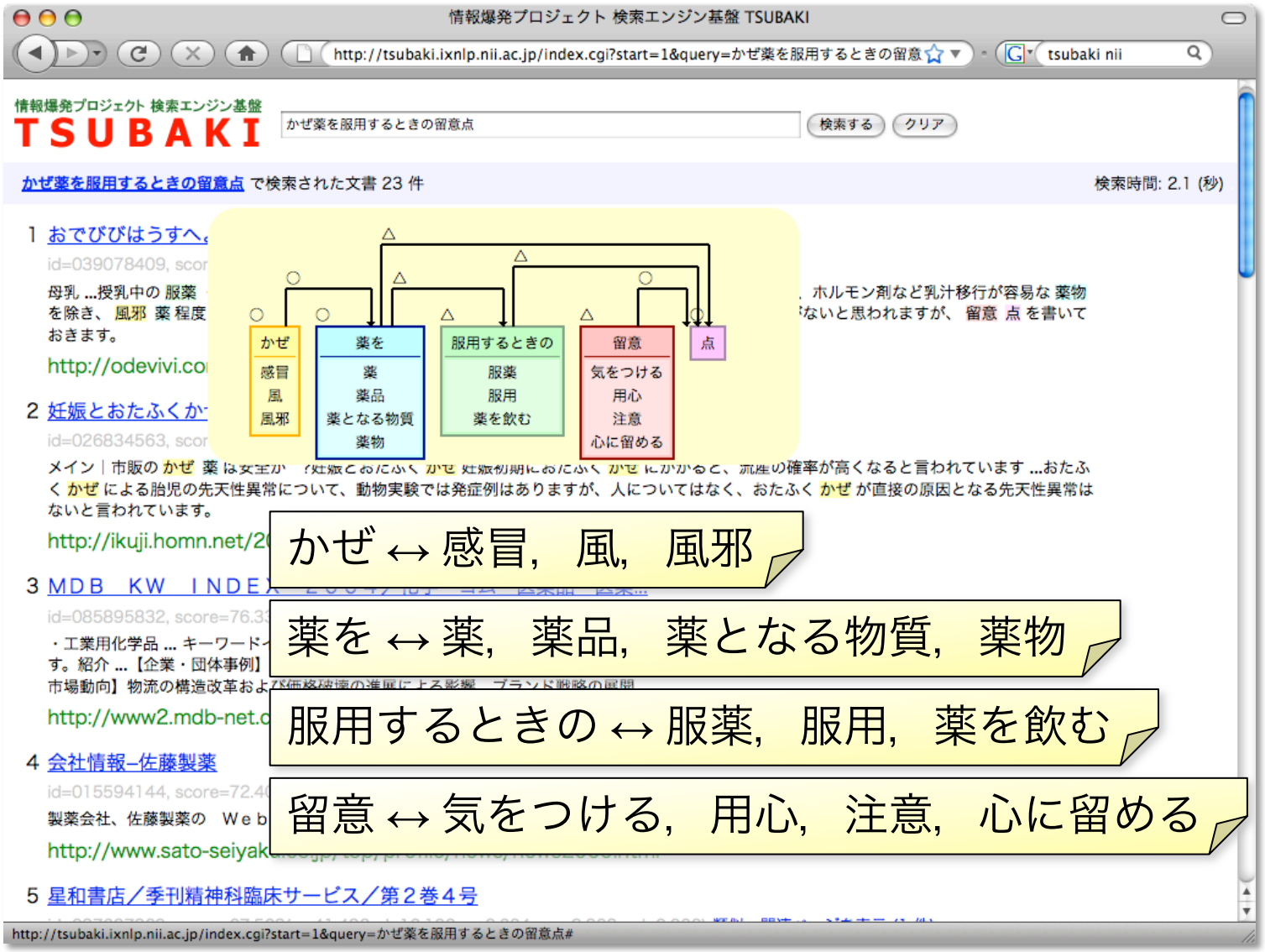
# 人間向け言い換え認識・生成

- e.g., 複数文書要約 [Barzilay 03]



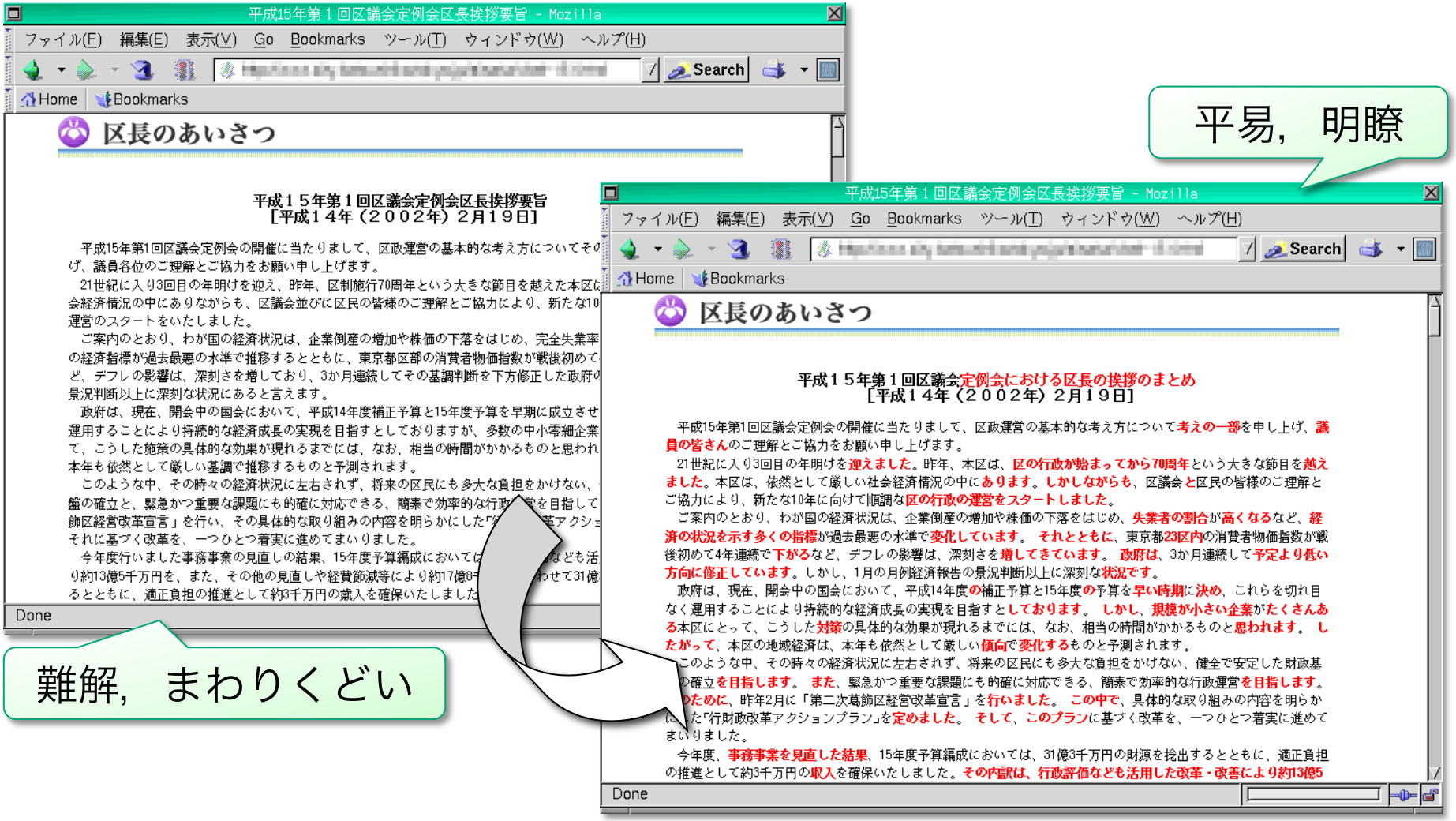
# 人間向け言い換え認識

- e.g., 情報検索 [Shibata+ 08]



# 人間向け言い換え生成

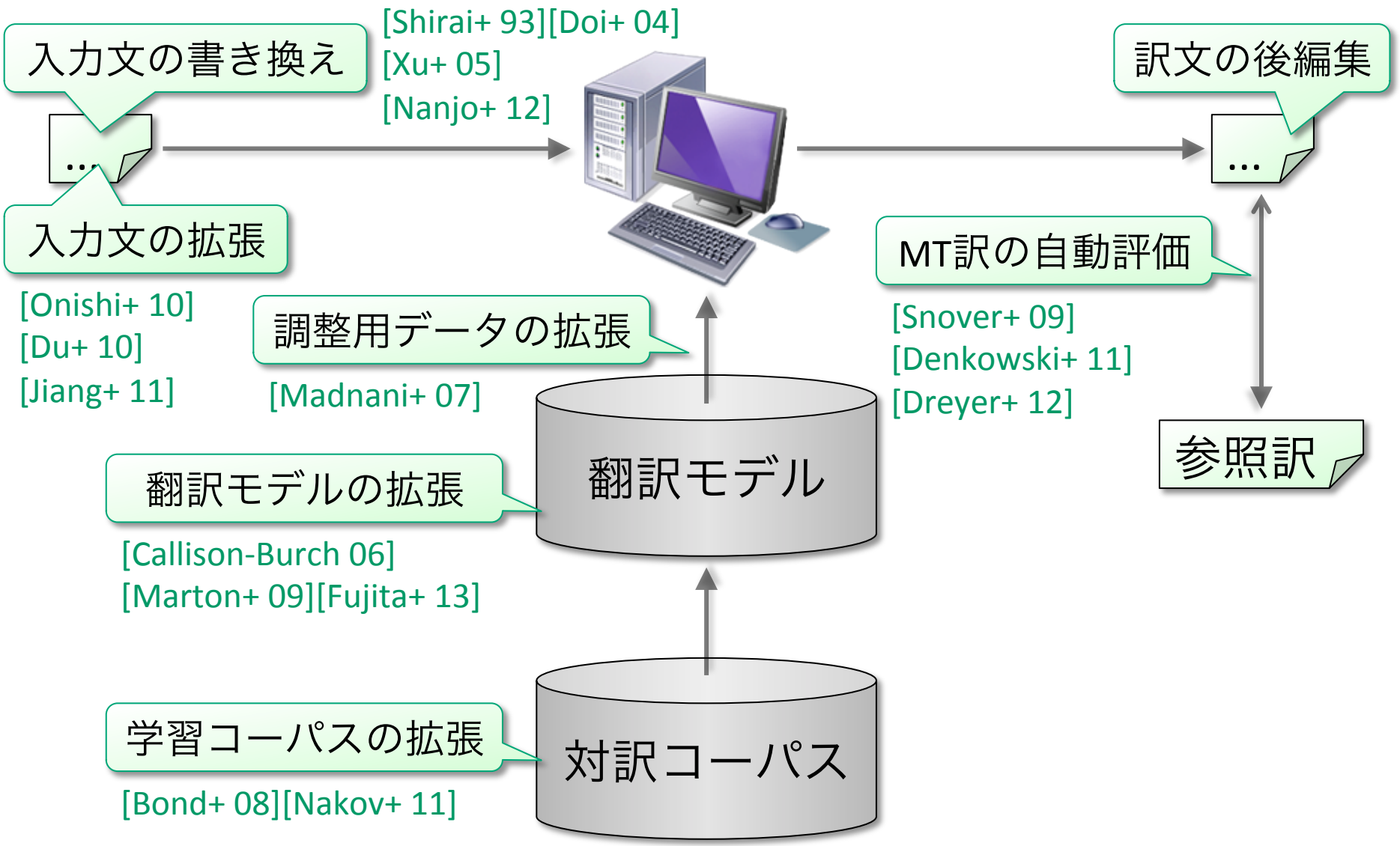
- e.g., ろうあ者向けのテキスト簡単化 [Inui+ 03]



# テキスト簡単化 ≠ 言い換え

- 言い換え (のサブセット) である必要はない
  - 伝達内容の省略
    - 言語表現の難解さは制御できても
    - 難解な概念はどうしても伝えられない → 抽象化，省略
  - 情報の構造化・可視化
    - e.g., 箇条書き，文字装飾，図示
- 用途に合わせた簡単化
  - 制限言語 [Huijsen 98]: 翻訳しやすい
  - 産業日本語 [熊野+ 09]: 人間・計算器の解釈が容易
  - やさしい日本語 [庵+ 13]: 日本語非母語話者でも分かる
    - e.g., NHK “News Web Easy” [田中+ 13]
  - 小学生向け語彙平易化 [梶原+ 15]

# MTへの各種応用


入力文の書き換え
[Shirai+ 93][Doi+ 04]
[Xu+ 05]
[Nanjo+ 12]
訳文の後編集
入力文の拡張
[Onishi+ 10]
[Du+ 10]
[Jiang+ 11]
調整用データの拡張
[Madnani+ 07]
MT訳の自動評価
[Snover+ 09]
[Denkowski+ 11]
[Dreyer+ 12]
翻訳モデル
参照訳
翻訳モデルの拡張
[Callison-Burch 06]
[Marton+ 09][Fujita+ 13]
学習コーパスの拡張
[Bond+ 08][Nakov+ 11]
対訳コーパス

# 応用を考える際の要検討事項

- 全自動か人間の支援という位置付けか
- どの種類の言い換えが必要か?
  - i.e., どのような違いを吸収したいか?
  - 言外の意味
    - 主題・焦点
    - 様式 (e.g., 形式性，丁寧さ)
    - 感情・態度
  - 指示的意味
    - テンス，モダリティ，述語や項の具体性
    - cf. 語用論的意味，参照先の同一性
    - cf. 含意，推論

# もくじ

(5) pp.85-91 (7)

(L5) 応用技術への導入
機械翻訳，評判情報処理，質問応答，文書要約，簡単化

(2) pp.34-52 (19)
(3) pp.54-65 (12)

(L4) 汎用的言い換え計算機構
言い換え認識，言い換え生成，評価・誤り分析

(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，言い換え知識ベース・コーパスの構築

(4) pp.67-83 (17)

(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(L1) 分類体系
定義と境界，現象の網羅・類型化，多言語対照分析

(1) pp.17-32 (16)
(6) pp.93-112 (20)

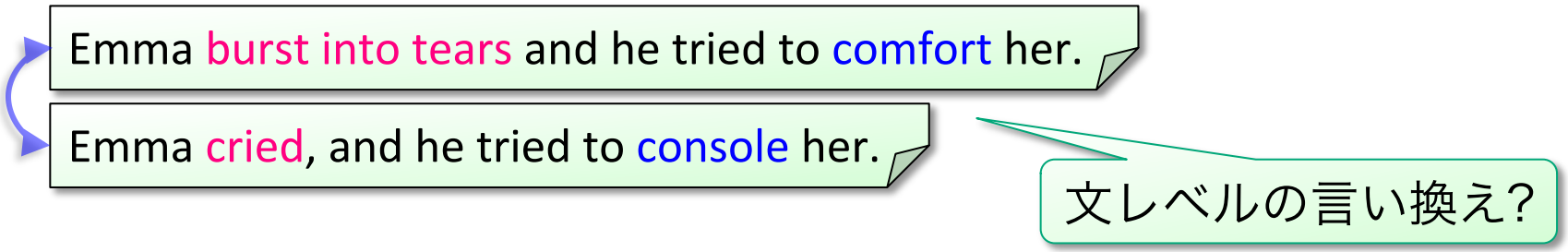
# 言い換えの分類

- 分類の意義
  - 人間が行っている計算の種類の俯瞰
  - 自動化の際の道標

- 既存の分類(等)
  - Phrase-level, Sentence-level, Discourse-level [IWP 2005]
  - 関係を説明する言語学要素に基づく3階層21種類 [Vila+ 11]
  - 自動獲得した表現対を8種類に分類 [Max+ 12]
  - Lexical, Phrasal, Syntactic [Ganitkevitch+ 13]
  - 語彙的な観点での分類 [Bhagat+ 13]
  - 大まかな6分類 [Lan+ 17]
  - 2階層25分類をMSRPにアノテーション [Kovatchev+ 18]

# 言い換えの分類における課題

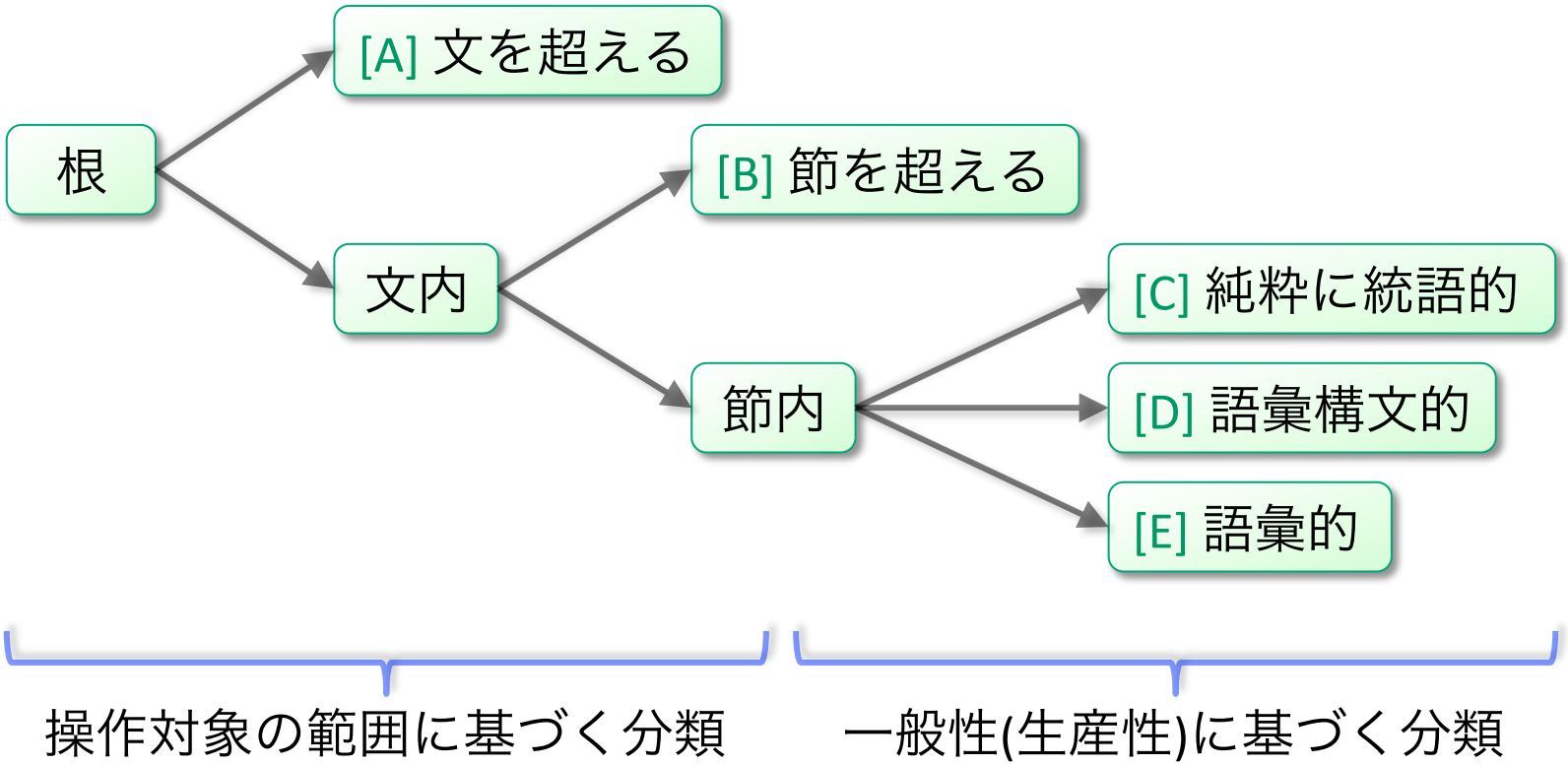
- 原子性




- 網羅性
  - 自動獲得結果の分析
  - 内省もフル活用
- 分類基準の客観化
  - 研究者間での会話の道具(メタ言語)とするため
  - e.g., 見かけ以上の素性の付与

# 我々の分類体系

- v3 5種類の大分類と60種類の小分類 [Fujita 10]
  - v1 乾解説論文 [乾+ 04]
  - v2 『言い換えのあれこれ』 http://paraphrasing.org/
  - v4 科研費『多言語対照版』をエンドレスにまとめ中



# [A] 文を超える言い換え

- 複数の節の分割・統合・再構成

スウェーデンの首都ストックホルムから南西部に位置するスモーランド地方が別名「ガラスの王国」とも呼ばれているのは、この地方にある2つの大きな町カルマルからベクショーにかけて、16ものガラス工場が点在しているからだ。

スモーランド地方は、スウェーデンのストックホルムから南西部に位置する。この地方は別名「ガラスの王国」とも呼ばれている。それは、この地方にある2つの大きな町カルマルからベクショーにかけて、16ものガラス工場が点在しているからだ。

# [B] 節を超える言い換え

- 分裂文 ↔ 無標文
  - 今週当選したのは、奈良県の男性でした。
  - 今週は、奈良県の男性が当選しました。
- 否定表現の移動
  - 返信しなければ、申込は取り消されます。
  - 返信すれば、申込は取り消されません。
- 節による連体修飾 ↔ ノ格による連体修飾
  - 目的地は屋根が赤い建物です。
  - 目的地は赤い屋根の建物です。
- 埋め込み節 ↔ 並列節

# [C] 純粋に統語的な言い換え

- 取り立て助詞の移動

ご飯は食べずに、辛いおかずばかりを食べていた。

ご飯は食べずに、辛いおかずを食べてばかりいた。

- 否定表現の言い換え

彼は有名な酒しか飲まなかった。

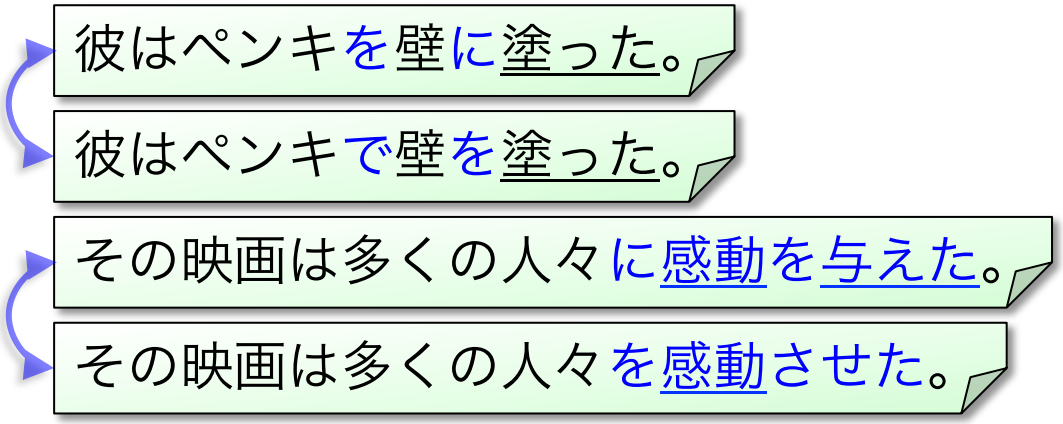
彼が飲んだのは有名な酒だけだった。

# [E] 語彙的な言い換え

- 一般化不可能な(つまり完全に語彙的な)言い換え
  → 網羅的な収集が不可欠
  - 内容語句
    - 太平洋を一望する桂浜公園の丘陵に建設された。
    - 太平洋を見渡せる桂浜公園の高台に建てられた。
  - 慣用表現
    - 岡山市は「ひかり都市」として脚光を浴びることとなりました。
    - 岡山市は「ひかり都市」として注目されることとなりました。
  - 機能表現
    - 秋には紅葉を見に多くの人が訪れるそうだ。
    - 秋には紅葉を見に多くの人が訪れるらしい。

# [D] 語彙構文的な言い換え

- 統語構造を変える言い換えではあるが
  - 語彙依存性がある




  - 語彙意味的な情報が不可欠
    - 細粒度の述語項構造情報
      - e.g., 意味役割，選択制限/選好，特質構造
    - 語間の関係に関する情報
      - e.g., 語彙的派生関係，反義関係

# [D]の一種: 態の交替

- ヲ → ガ の受動態
  - 先生は(ガ)その学生を褒めた。
  - その学生は(ガ)先生に褒められた。
- ニ → ガ の受動態
  - 彼女は(ガ)彼に指輪をねだった。
  - 彼は(ガ)彼女に指輪をねだられた。
- 間接受身
  - 誰かが彼の自転車を盗んだ。
  - 彼は(ガ)誰かに自転車を盗まれた。
- ガ → ヲ の使役
  - 彼は(ガ)心労でやつれた。
  - 心労が彼をやつれさせた。
- ガ → ニ の使役
  - 彼は(ガ)その失敗で自信をなくした。
  - その失敗が彼に自信をなくさせた。

# [D]の一種: 動詞交替

- 自他交替
  - 傘が雨を濡らす。
  - 傘で雨が濡れる。
- 自他交替・再帰
  - 円のレートが下がった。
  - 円がレートを下げた。
  - フランスの領地が拡大した。
  - フランスが領地を拡大した。
- 相互格の交替
  - その車が一台の自転車と衝突した。
  - その車と一台の自転車が衝突した。
- 授受の交替
  - 健は悟から車を買った/借りた。
  - 悟は健に車を売った/貸した。
- 場所格交替
  - 彼はペンキを壁に塗った。
  - 彼はペンキで壁を塗った。
- 湧出動詞の交替
  - 柳が芽をふく。
  - 柳に芽がふく。

# [D]の一種: 品詞交替/範疇交替

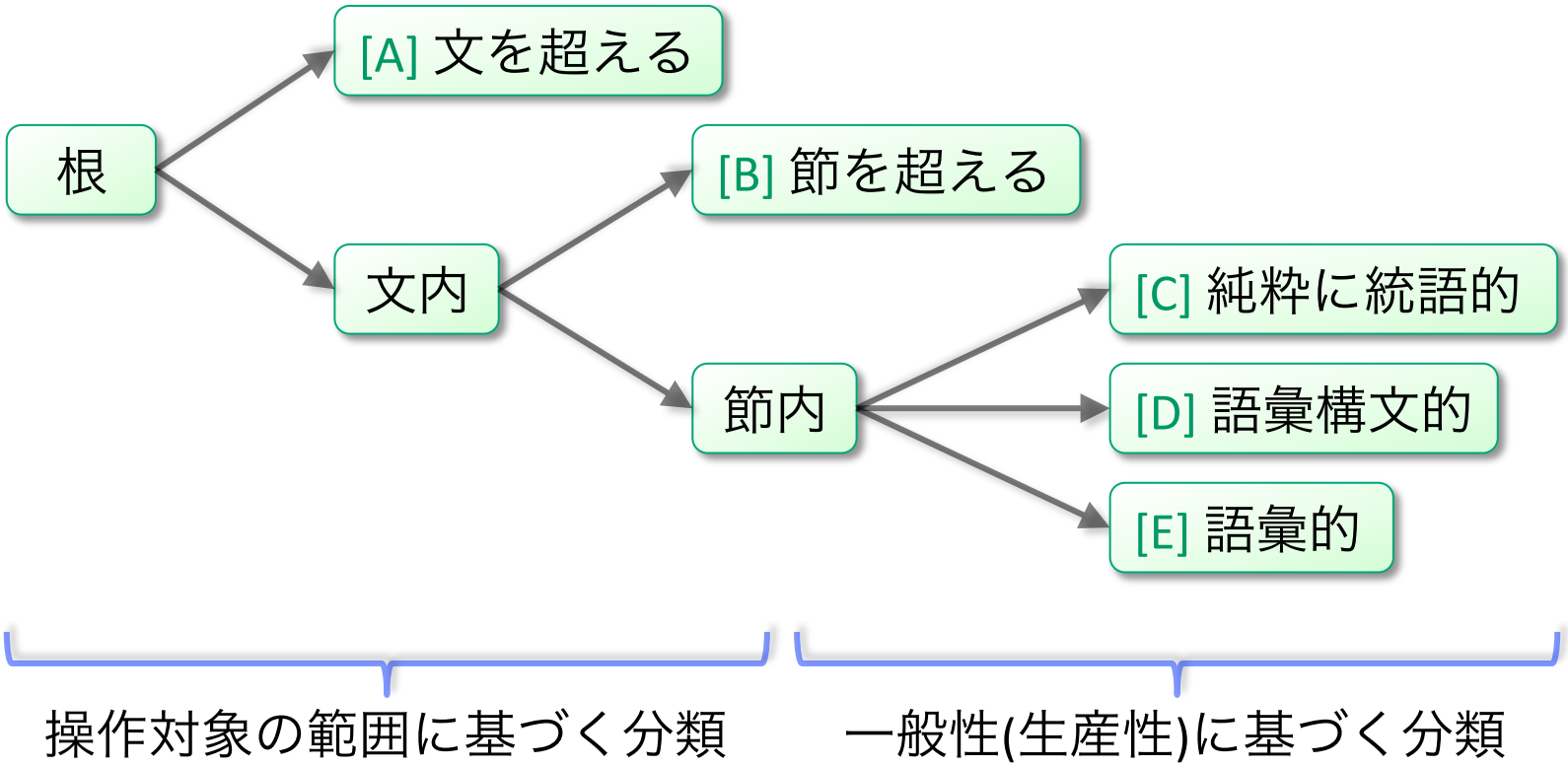
- 機能動詞構文 (名詞 ↔ 動詞)
  - その映画は多くの人々に感動を与えた。
  - その映画は多くの人々を感動させた。
- 形容詞 ↔ 動詞
  - 風が強くなってきたようだ。
  - 風が強まってきたようだ。
- 名詞 ↔ 形容詞
  - 日焼けのため肌が赤みを帯びている。
  - 日焼けのため肌が赤くなっている。

# [D]の一種: 構造変換

- 節と句の主辞交替
  - 仕事がとても魅力的だ。
  - とても魅力的な仕事だ。
- 名詞句の主辞交替
  - 処理の効率化が必要だ。
  - 効率的な処理が必要だ。
- 格と副詞の主辞交替
  - 英語の教育を徹底する。
  - 英語を徹底的に教育する。
- 連体修飾 ↔ 連用修飾
  - 有意義な週末を<br>お過ごしください。
  - 週末を有意義に<br>お過ごしください。
  - 1件の開示請求があった。
  - 開示請求が1件あった。
- 連体修飾 ↔ 格
  - 彼の顔が真っ赤だ。
  - 彼は顔が真っ赤だ。

# まとめ

- 5種類の大分類と60種類の小分類 [Fujita 10]
  - ポスト [乾+ 04] [藤田 07]
  - 原子性，(網羅性)，操作対象の範囲，一般性(生産性)
  - TODO: 必要な言語資源，影響の及ぶ範囲，差分の分類



# 議論 (1/3): 分類体系の妥当性

- 文というのは形式的な単位に過ぎない
  - 「節を超える」と「節内」の区別の方が単純
- [E] 語彙的な言い換えに押し込めるのがイマイチな例
  - 個々の語に依存するが生産性は高くない言い換え
    - Bill flew across the ocean.
    - Bill crossed the ocean by plane.
  - ドメイン知識が不可欠な言い換え
    - ローマは連敗を2で止めた。
    - ローマは3試合ぶりに勝点を得た。
    - L. メッシは先日のR・マドリード戦で無得点に終わった。
    - L. メッシは先日のR・マドリード戦でゴールをあげられなかった。

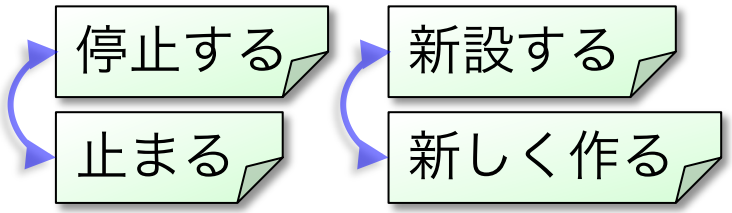
# 議論 (2/3): 各々の実現方法

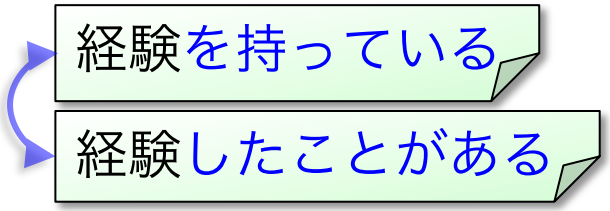
- [A-C] 種類は少ない → 過剰生成を許容しつつ書く
- [D] 統語と意味の相互作用を要解明 → 語彙意味論
  - Meaning-Text Theory [Mel'čuk+ 87]
  - 語彙概念構造 [Jackendoff 91]
  - 生成語彙論 [Pustejovsky 95]
  - FrameNet [Baker+ 98]
  - Verb classification [Levin 93]
    → VerbNet [Kipper+ 00]
- [E] 要丸覚え → 自動獲得に頼るしかない
  - 既存の語彙資源も最大限に活用
    - e.g., シソーラス，辞書の定義文

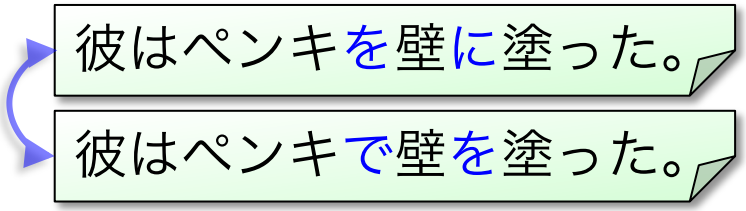
大規模な言い換え知識獲得によって，[E]と同じように扱える見込みが出てきた．

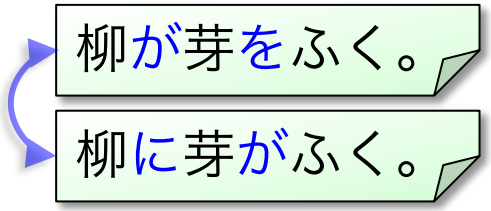
# 議論 (3/3): 現象を網羅することの難しさ

- 内省の限界
- ロングテール性
  - 頻出:
    - e.g., 漢語動詞→和語動詞+α
  - 中程度:
    - e.g., 機能動詞構文→単一の動詞+α
  - 更にレア:
    - e.g., 与格交替, 場所格交替
  - 極めてレア:
    - e.g., 源泉の交替



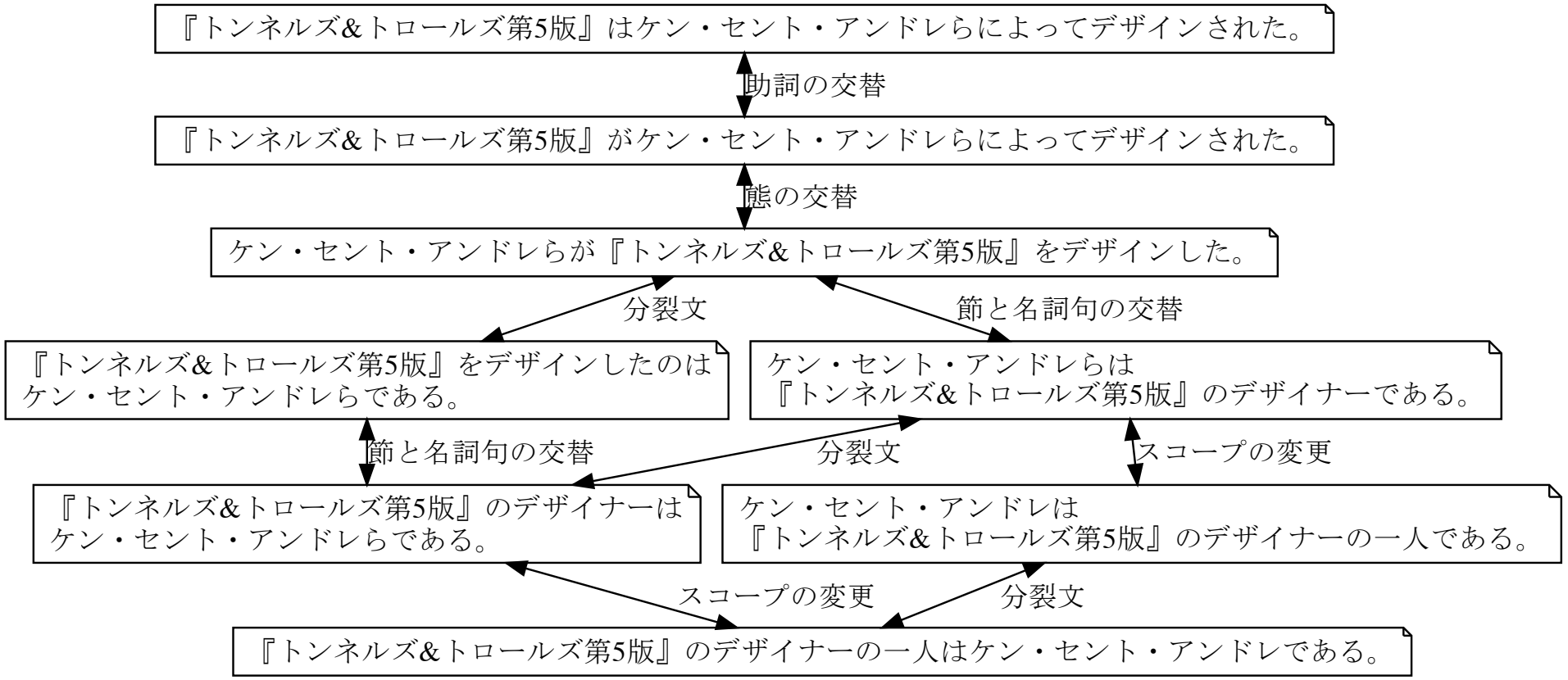



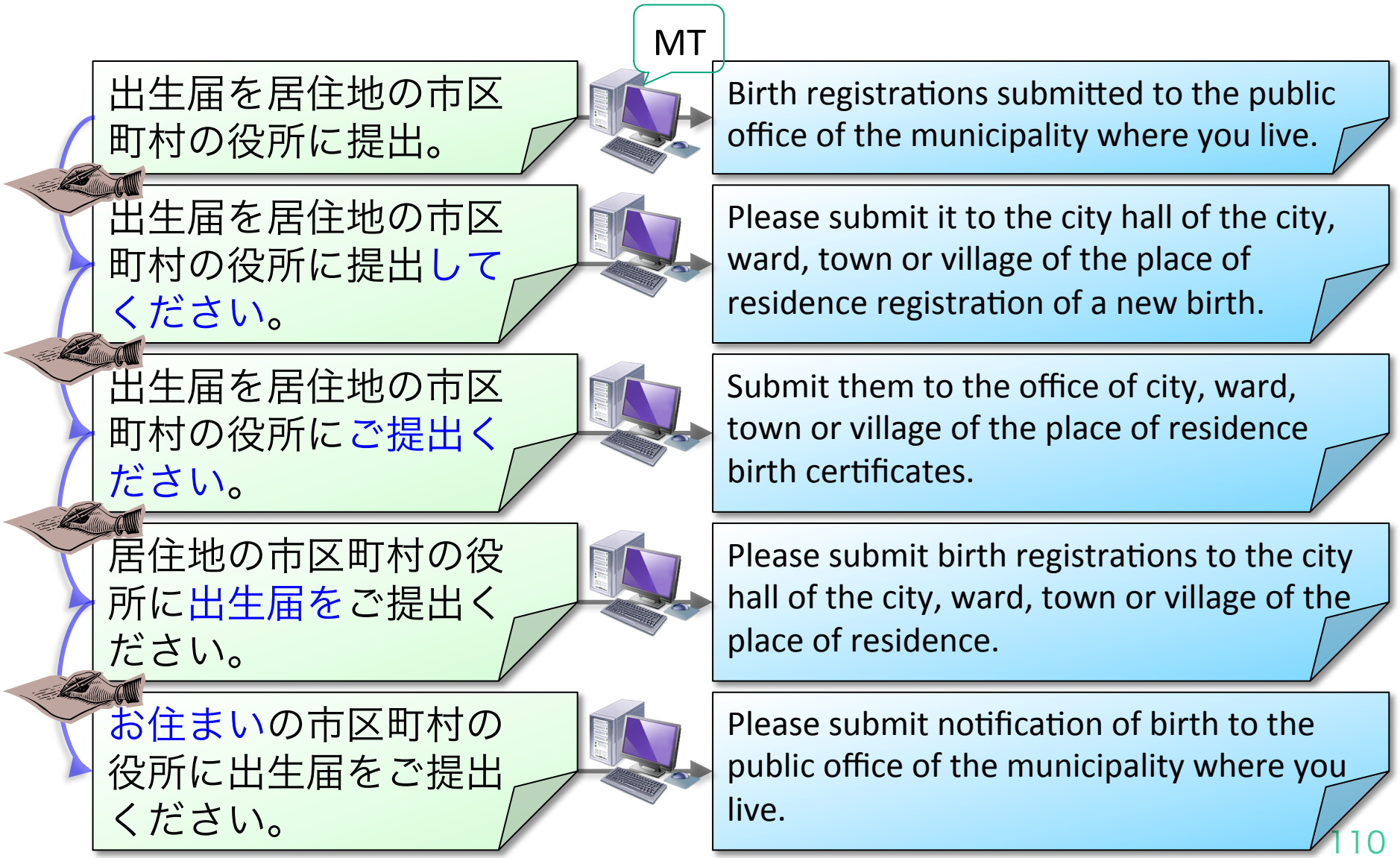
# 我々の試み (1/2)

- 既存のデータ中の事例の分解・分類 [藤田+ 15]
  - 元データ収集時のバイアスの問題
  - Twitter URLのアプローチは有用そう [Lan+ 17]



# 我々の試み (2/2)

- MT向け前編集を通じた事例収集 [宮田+ 15;16;17]



# 収集した前編集事例

- 4データセット(各100文) → NICTのGPMT
  - 12,867事例
  - MTのポテンシャル: 90%超の文に対して良質な翻訳を達成
  - 病院内会話ドメインは元々得意

| データセット | 平均文長 (標準偏差) | 書き換え数 | | | | 原文の数 | |
|---|---|---|---|---|---|---|---|
| | | Total | Avg. | Med. | Max | MT可能 | 原文まま |
| 病院 | 12.1 (4.5) | 1199 | 12.0 | 3 | 105 | 97 | 40 |
| 自治体 | 21.3 (12.0) | 2119 | 21.2 | 14.5 | 89 | 97 | 3 |
| BCCWJ | 26.9 (16.0) | 3823 | 38.2 | 26.5 | 209 | 86 | 0 |
| Reuters | 34.8 (12.6) | 5546 | 55.5 | 45 | 258 | 93 | 4 |

# 事例の分解と書き換えの種類の分類

- 蓄えられた事例は必ずしもプリミティブではない
  → 人手で分解 → 2階層53種類 [宮田+ 17]

十日前後で登録のクレジットカードから引き落としを行います。

1. プリミティブな書き換えへの分解
2. 表層的な対応と差分の抽出
3. 書き換えの種類の類型化(大域的に)

～で～から～→～から～で～ **語順変更**

登録のクレジットカードから十日前後で引き落としを行います。

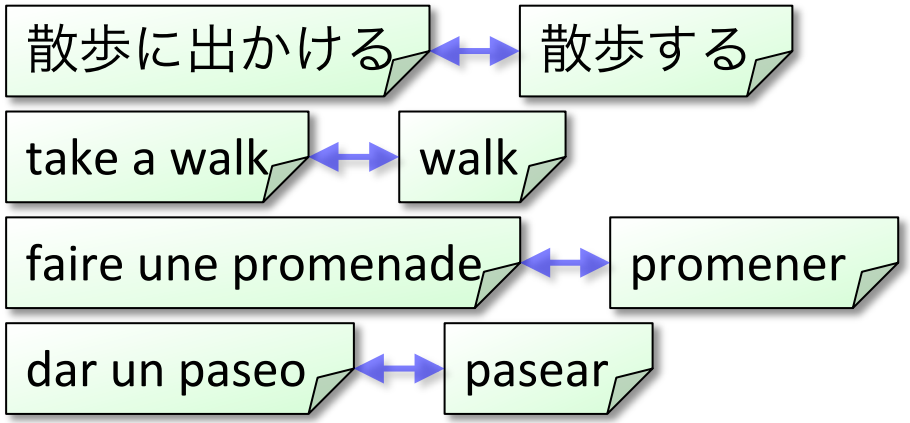
～を行います→～が行われます **態の変更**

登録のクレジットカードから十日前後で引き落としが行われます。

# 残っている研究課題
# (まとめに代えて)

# 研究課題: 工学的な観点から

- 知識開発: 様々な種類の言い換えのカバー
  - 知識の形式
  - 資源の集約
  - 効率的な評価手法の確立
- 計算機構の高度化
  - 言い換え元の表現の多義性解消
  - 評価用データセットの構築
  - 汎用化を目指すよりは目的も含めてNN化?
- 応用
  - 効果の多寡の調査 (oracle)
  - 言い換えの種類ごとの選択的な利用
    - 言い換え以外の問題も含めた全体の見通し

# 研究課題: 言語学的な観点から

- 包括的な分類体系の構築
  - 現象分析，知識獲得，応用のいずれの場面でも必要
  - 作り方や評価方法の検討
  - 言語横断的な現象の体系化と分析 (道半ば)