言い換え技術の研究動向 (Early 2016)

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

自己紹介

- 藤田 篤 (ふじた あつし)
- 研究分野
  - 自然言語表現の言い換え (FY1999～)
  - 機械翻訳・翻訳関連研究 (FY2003, FY2011～)
- 略歴
  - 2005年3月: NAIST 松本研 修了→ 京大-NTTCS研 → 名大 佐藤研 → 未来大 → NICT

NICT客員　NRC客員

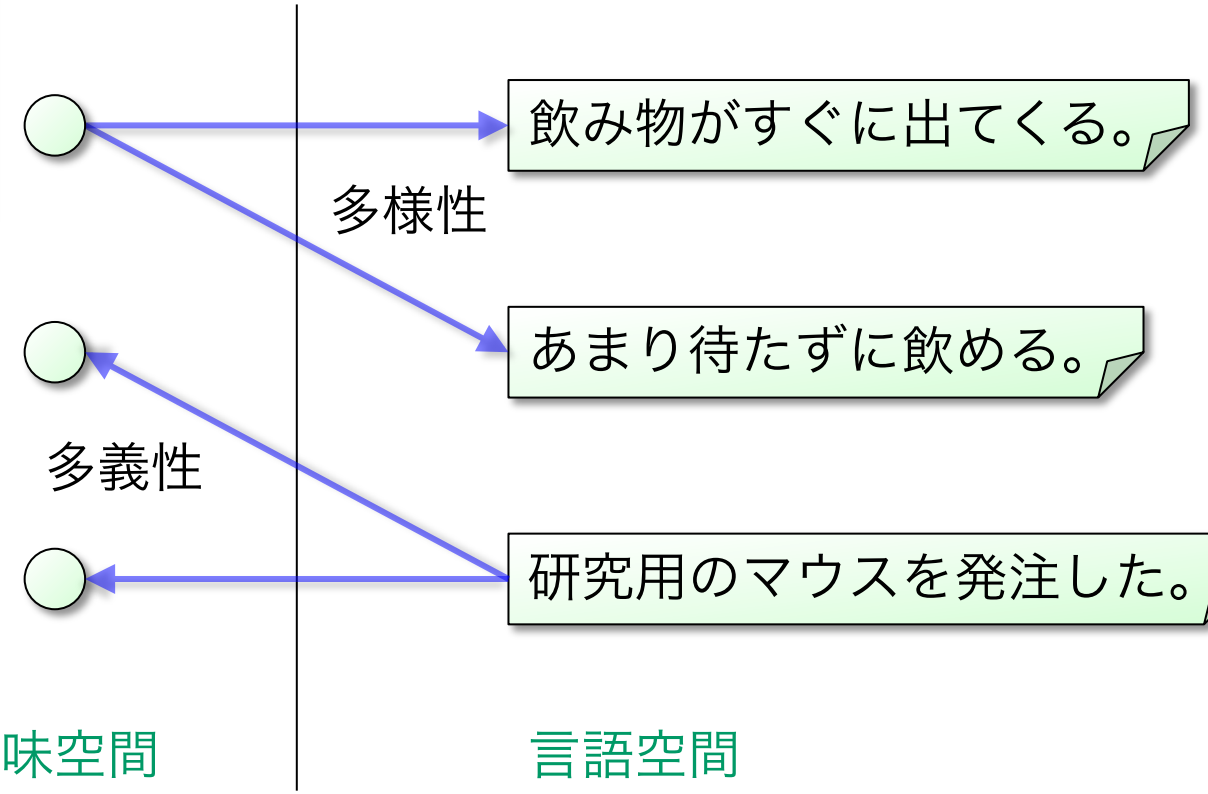
今日の話題: 言い換え

- 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









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
  - 個々の表現は別個のアルファベットではない

- 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

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，ベンチマーキング，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，データベース構築

(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，ベンチマーキング，誤り分析

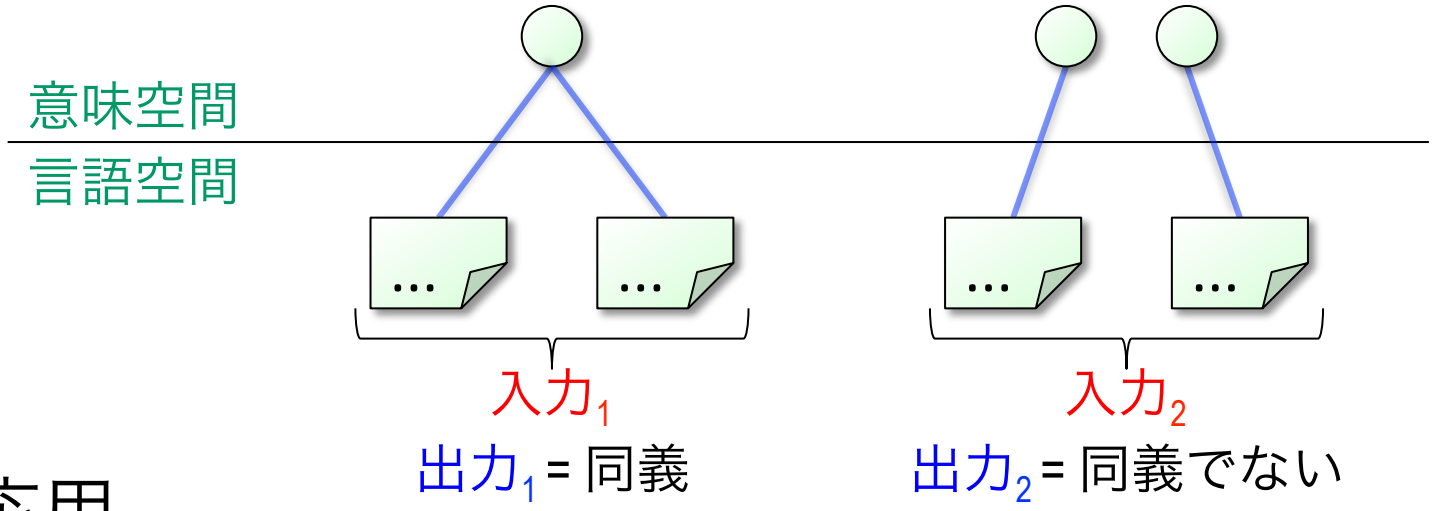
(L3) 言い換えの計算に必要な知識の整備
大規模知識獲得，データベース構築

(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)

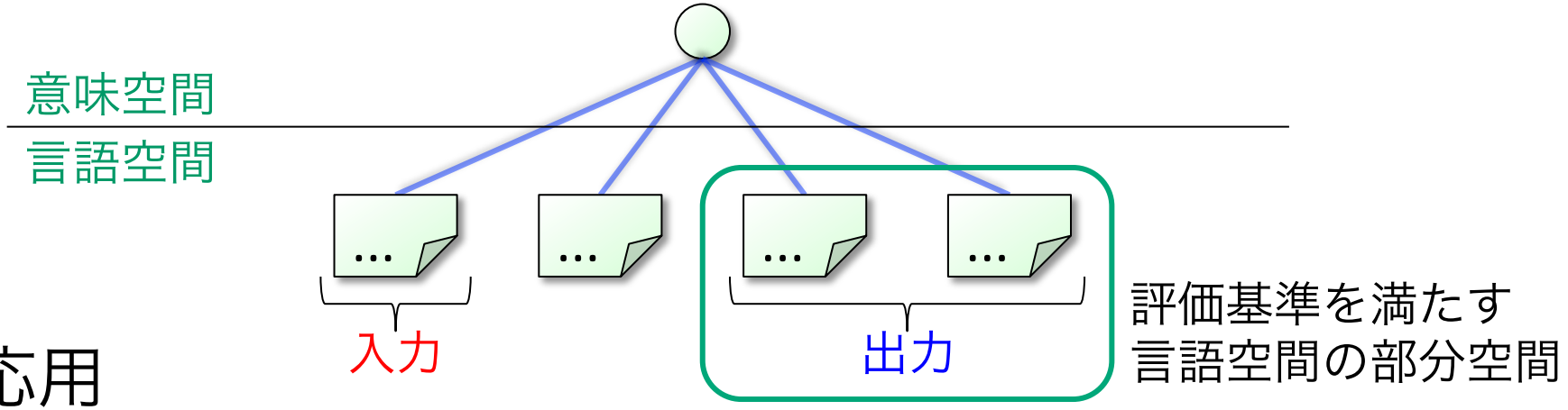
- 言い換え認識 (paraphrase recognition / identification)
  - 入力: 同一言語の複数(一般的に2つ)の異なる言語表現
  - 出力: 入力された言語表現が同義か否か (あるいは0〜1)



- 応用
  - 情報検索, 質問応答, 複数文書要約, 剽窃検出
    (実際には言い換え生成との併用)

(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
  - 簡単化，文圧縮，MTの前処理，TTSの前処理

(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
  - 表現間の関係に関する知識
- 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]

- 言い換え知識獲得 (paraphrase acquisition)
  - [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

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

(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 文章読解支援 TTS前処理 MT後編集 ステガノグラフィ	剽窃検出 翻訳メモリ検索 情報検索 質問応答 東口ボ!?
機械処理向け	SMTのモデル拡張 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

- 全体像の解明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，ベンチマーキング，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，データベース構築

(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
明示的意味が等しい

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

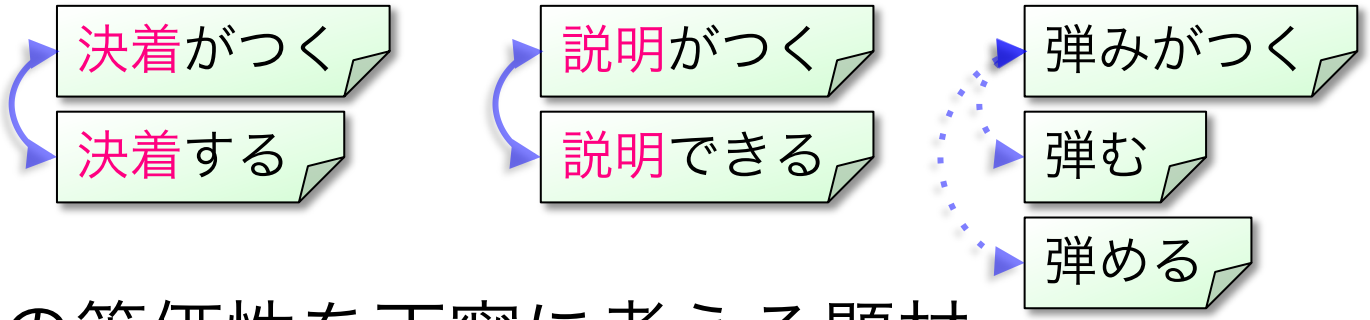
うちの子はトマトをよく食べる。
うちの子はトマトが好きだ。

常に真とは限らない

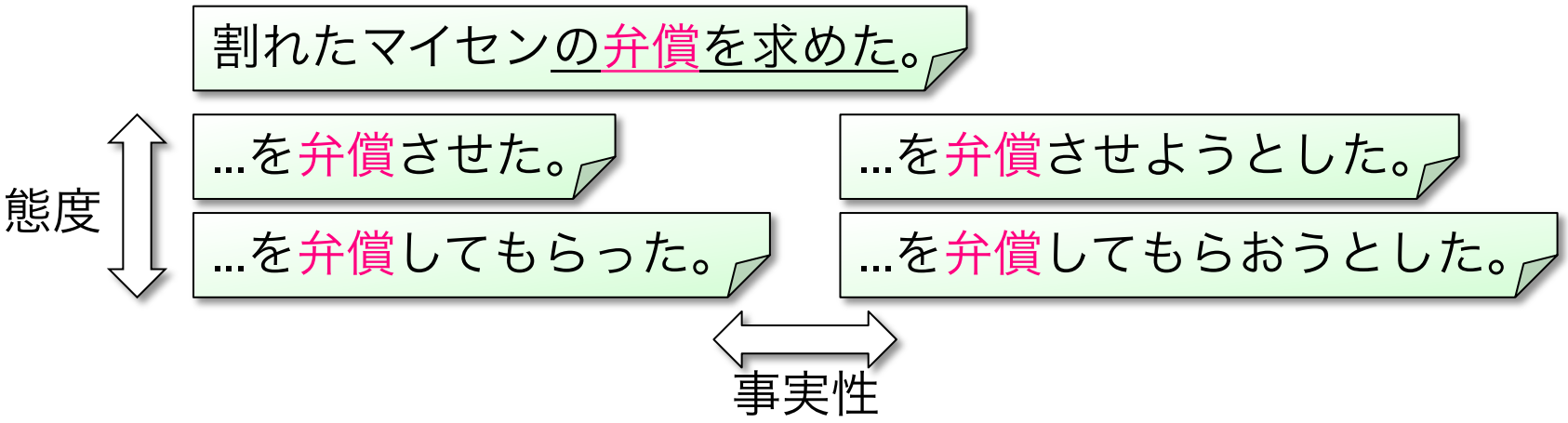
- 分類の意義
  - 人間が行っている計算の種類の俯瞰
  - 応用技術における取捨選択の判断材料

(L2) 言い換えの事例研究

- 特定の種類の言い換えの成立理由の解明
  - 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



今日のトークの趣旨

- 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
      - 評価における課題
    - 応用研究
  - (比較的)言語学的な研究
    - 現象の網羅と類型化
    - 事例研究
- (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.77-84 (8)

(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.33-44 (12)
(3) pp.46-56 (11)

(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，ベンチマーキング，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，データベース構築

(4) pp.58-75 (18)

(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.16-31 (16)
(6) pp.86-105 (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?

しのぎをけずる
激しく戦う

似ている
瓜二つだ

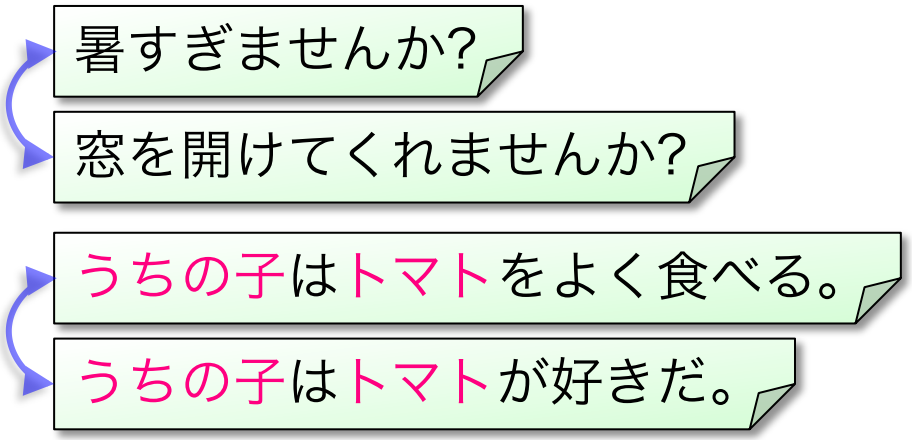
健は悟から車を買った。
悟は健に車を売った。

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。

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。

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

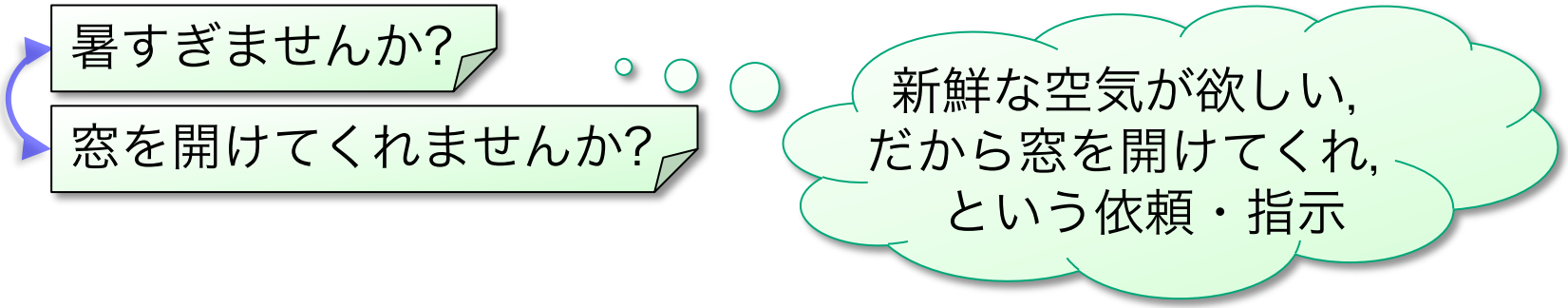
意味処理の例題としての言い換え

- 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

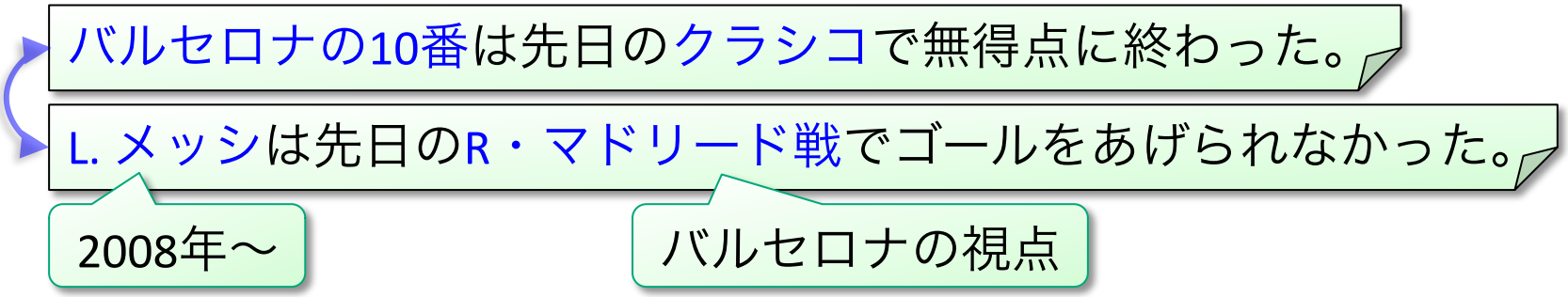
- コミュニケーションにおける発語の役割



- 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

b. 参照先が等しい

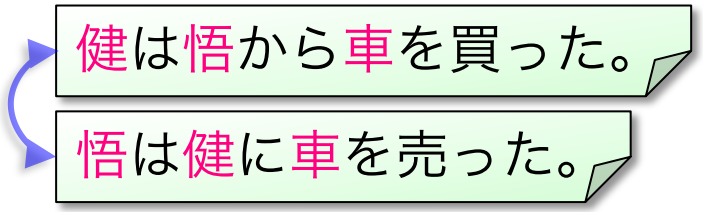
- 参照的意味 (referential meaning)



  - 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

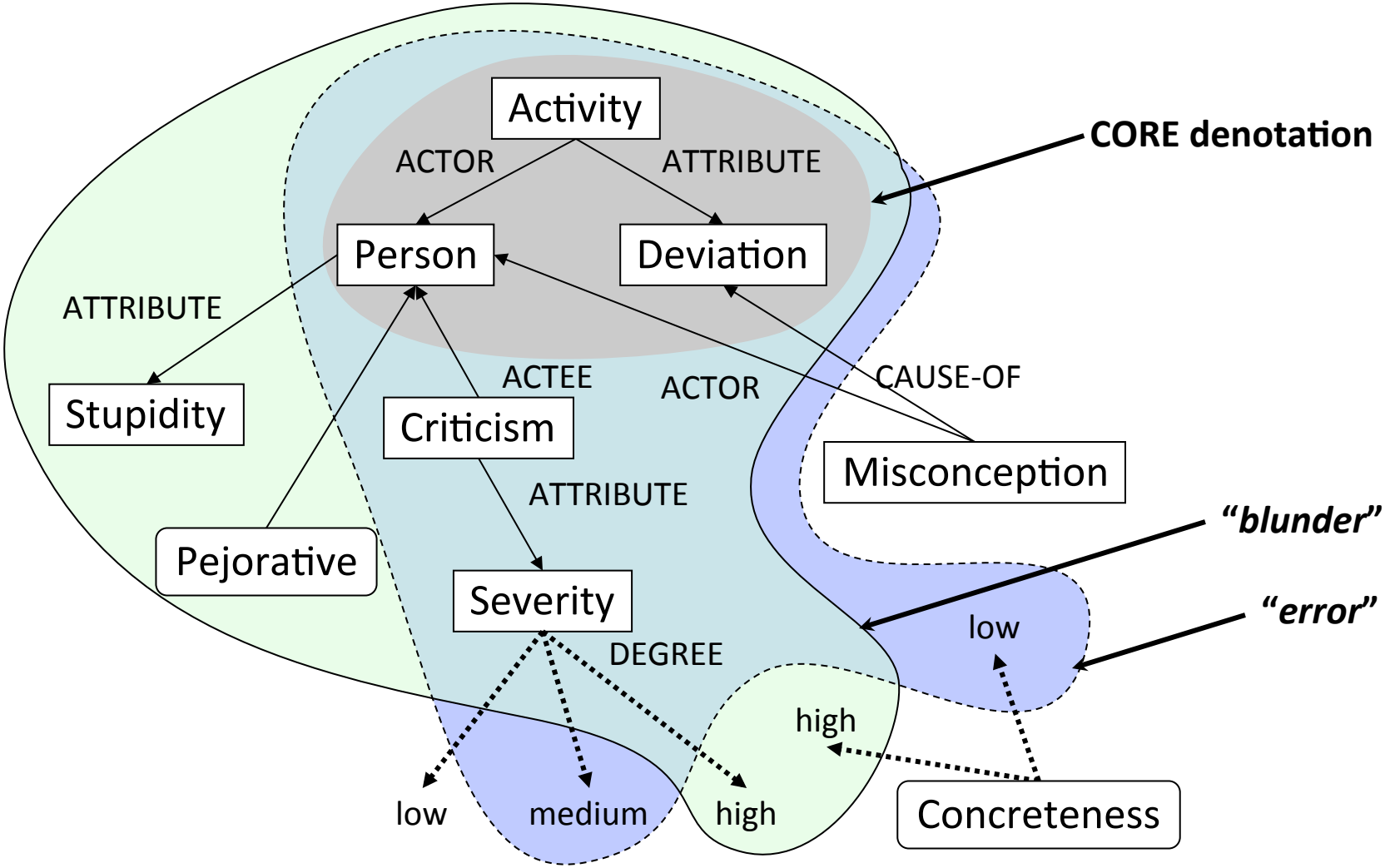
> paraphrase$_{(v)}$: to express in a shorter, clearer, or different way what someone has said or written

> 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
    - 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

今日の話題: 言い換え

- 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







飲み物がすぐに出てくる。

あまり待たずに飲める。

研究用のマウスを発注した。

多様性

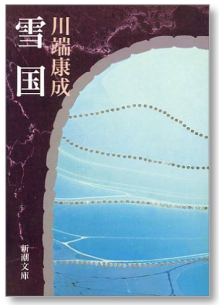
多義性

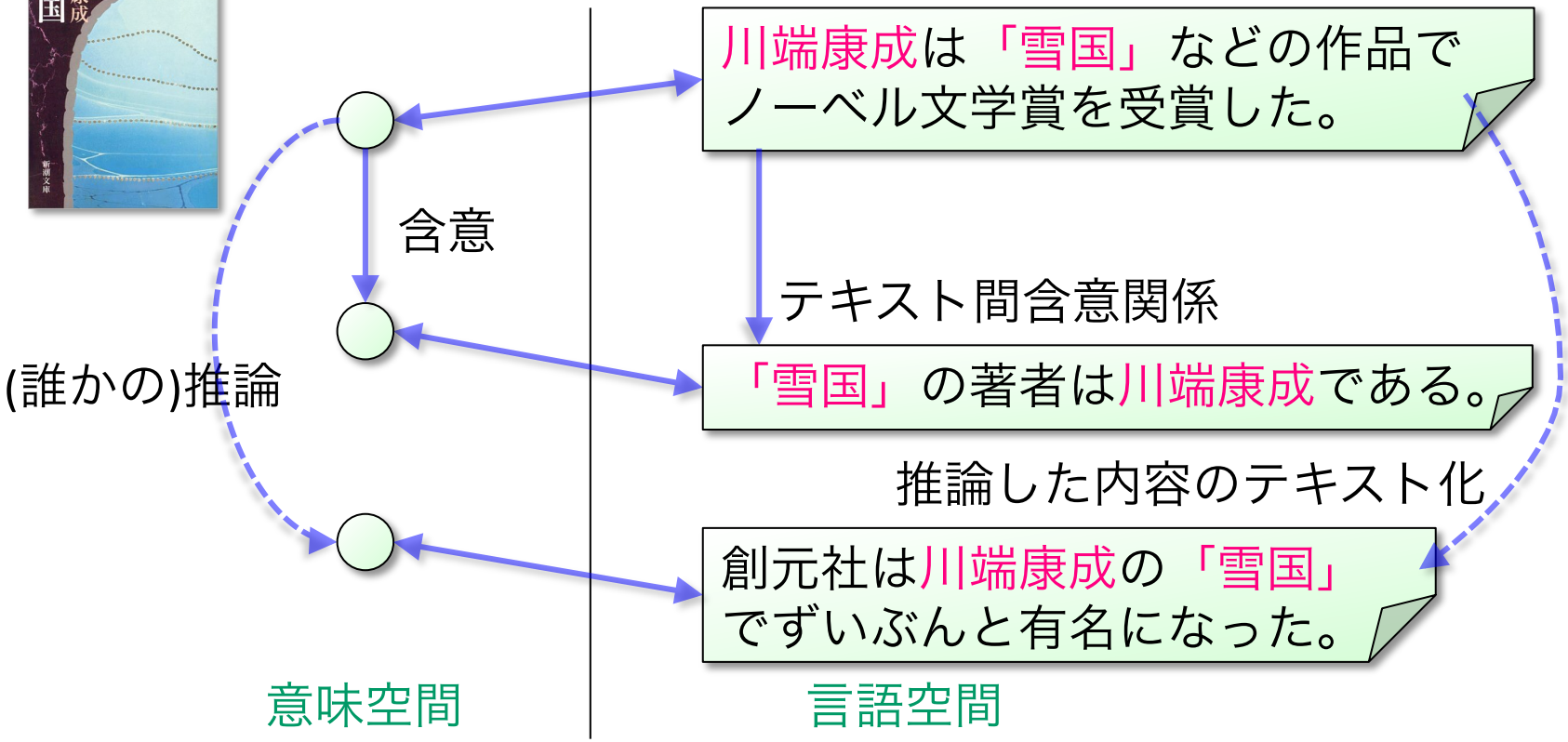
意味空間

言語空間

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



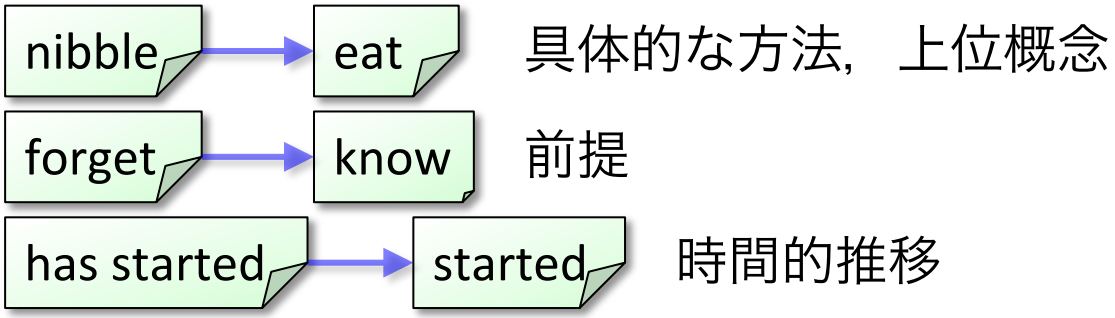


テキスト間含意関係 (textual entailment)

- 同じ意味とは限らない
  - X → Y: Y holds if X holds



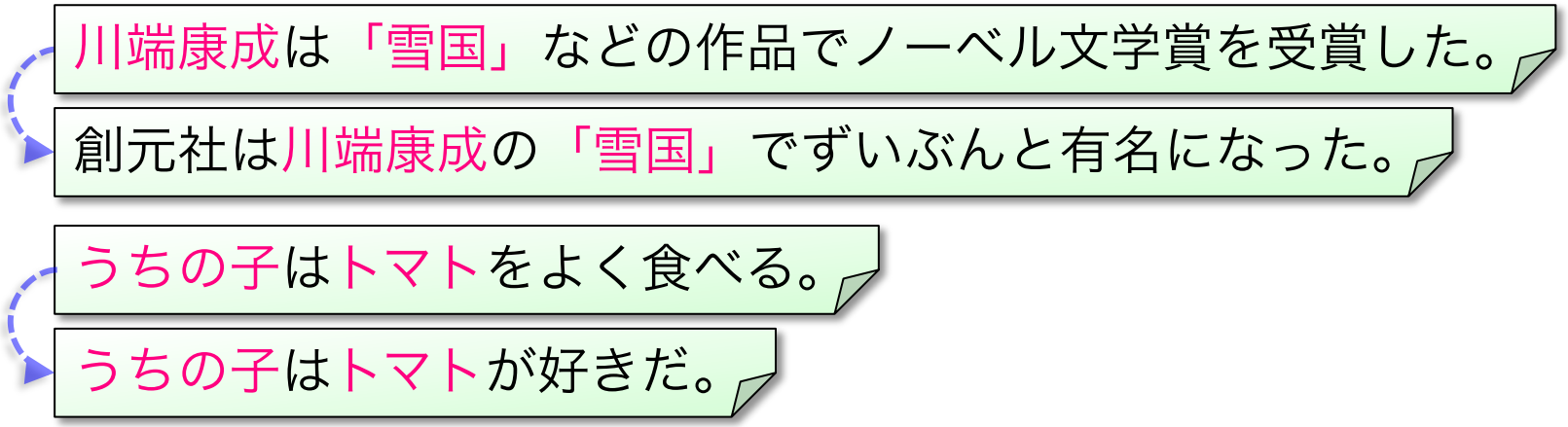
  - e.g., WordNet における語彙的含意 [Miller+, 85]



  - 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
  - 発語内行為とも異なる

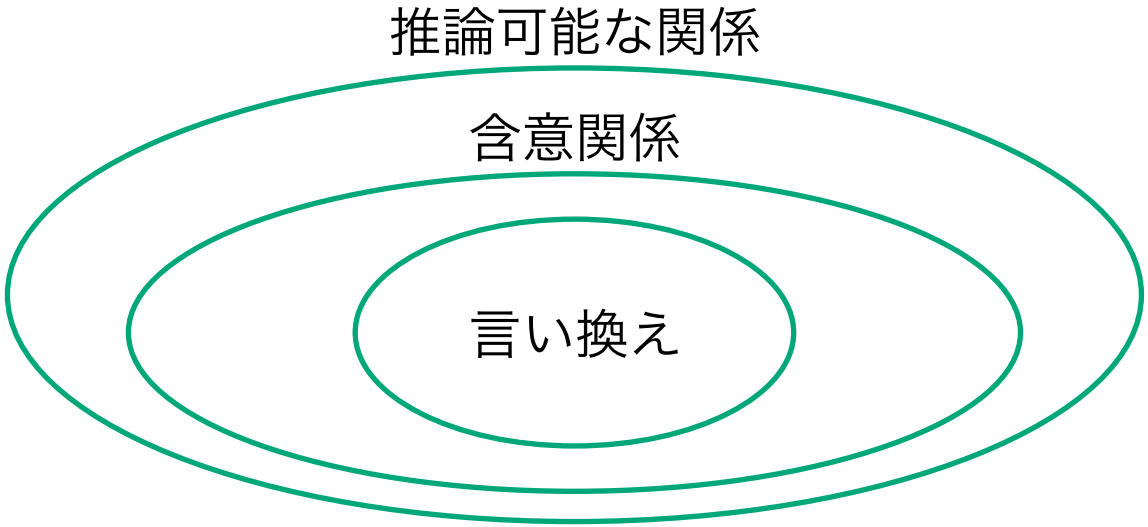


  - ただしこの計算が役立つ場合もある [Pantel+, 07]

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.77-84 (8)

(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.33-44 (12)
(3) pp.46-56 (11)

(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，ベンチマーキング，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，データベース構築

(4) pp.58-75 (18)

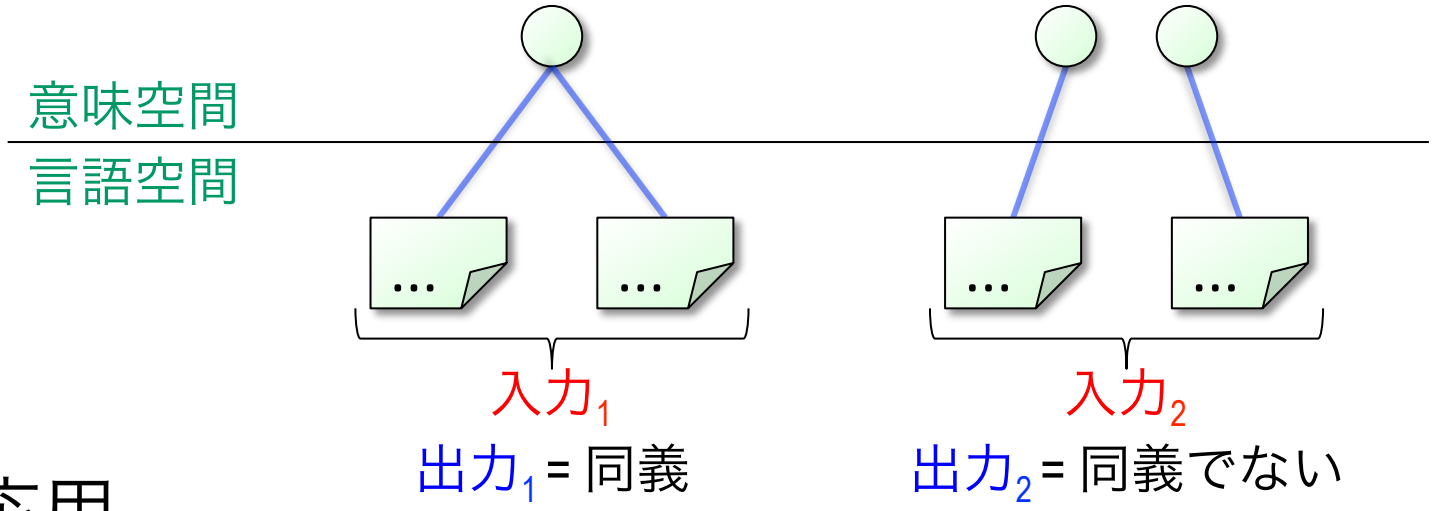
(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(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.16-31 (16)
(6) pp.86-105 (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 (あるいは0〜1)



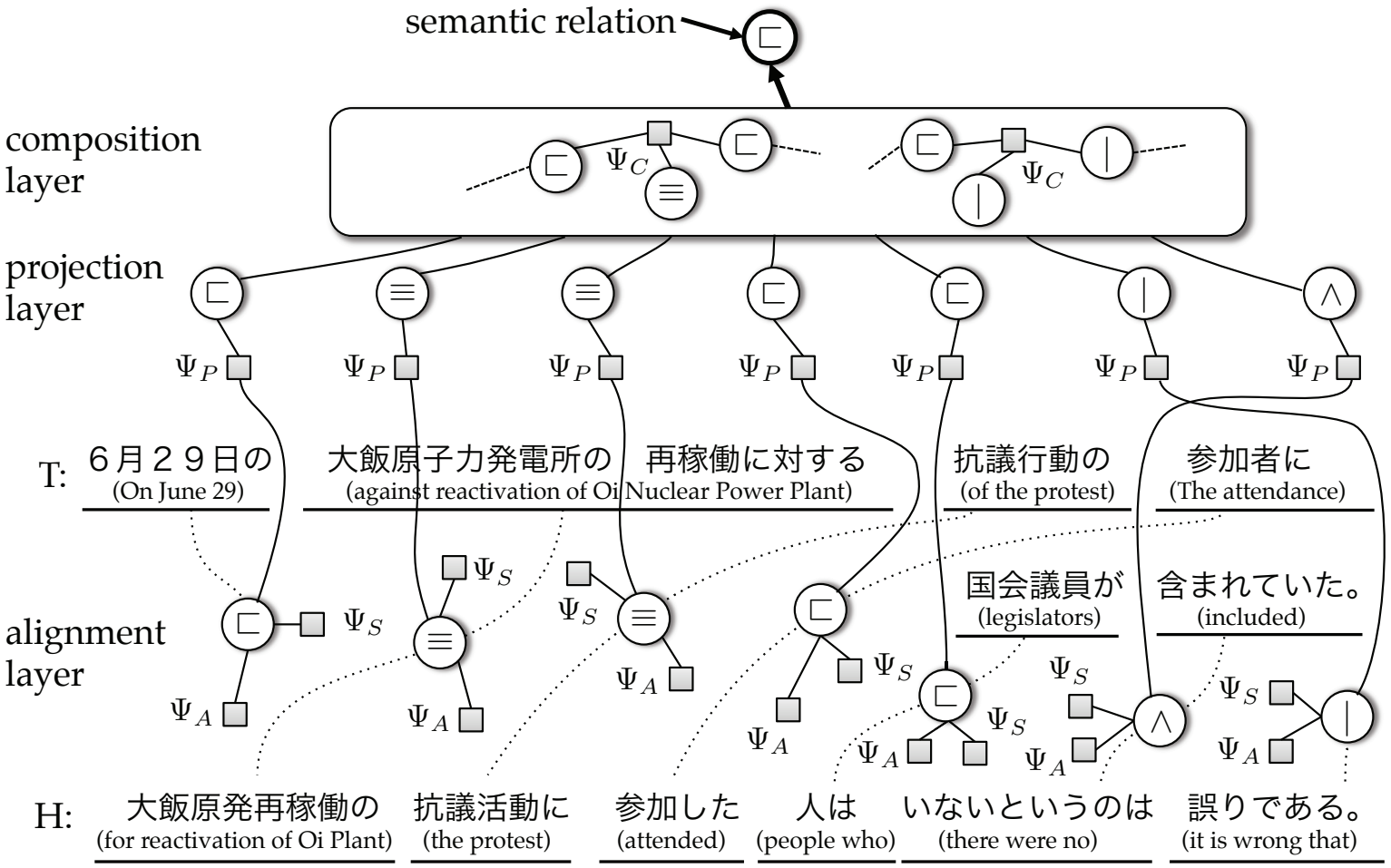
- 応用
  - 情報検索, 質問応答, 複数文書要約, 剽窃検出 (実際には言い換え生成との併用)
- 評価方法: 正解データを用意しておいて照合
  - e.g., Microsoft Research Paraphrase Corpus [Dolan+, 04]

2値分類問題/回帰問題としての定式化

- 機械学習の例題
  - 学習フェーズ
    - 学習データの各言語表現対における対応箇所と差分を抽出
    - 各特徴の重要度を正解ラベルに基づいて最適化
  - 判定フェーズ
    - 判定対象の言語表現対の特徴量を抽出
    - 学習フェーズで推定した重要度を考慮して同義性を判定
  - ベンチマーキング用データを使って比較
    - e.g., Microsoft Research Paraphrase Corpus (MSRP) [Dolan+, 04]
- 特殊ケース
  - 同義の文対の収集 [Dolan+, 04][Hashimoto+, 11][Xu+, 14]
  - 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例 (1/3)

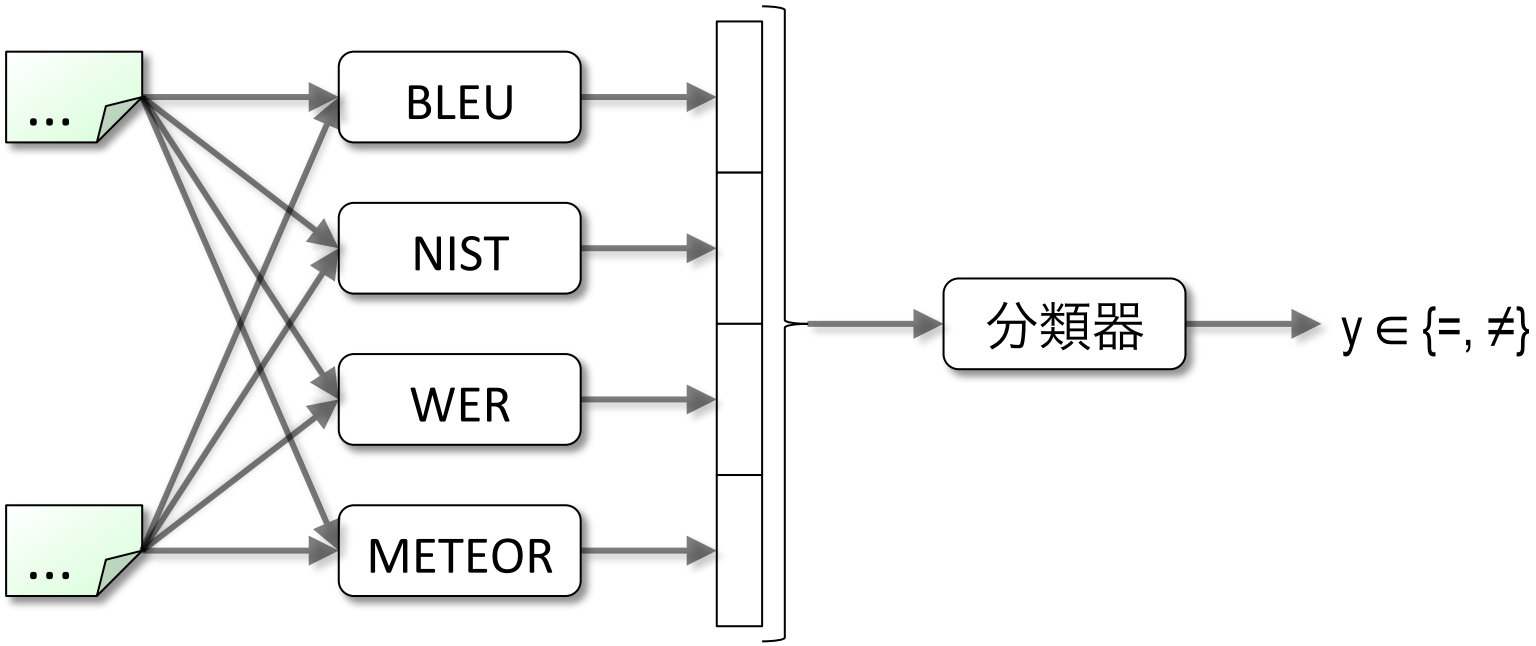
- 表層上の論理的対応関係の合成 [Watanabe+, 12]



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.

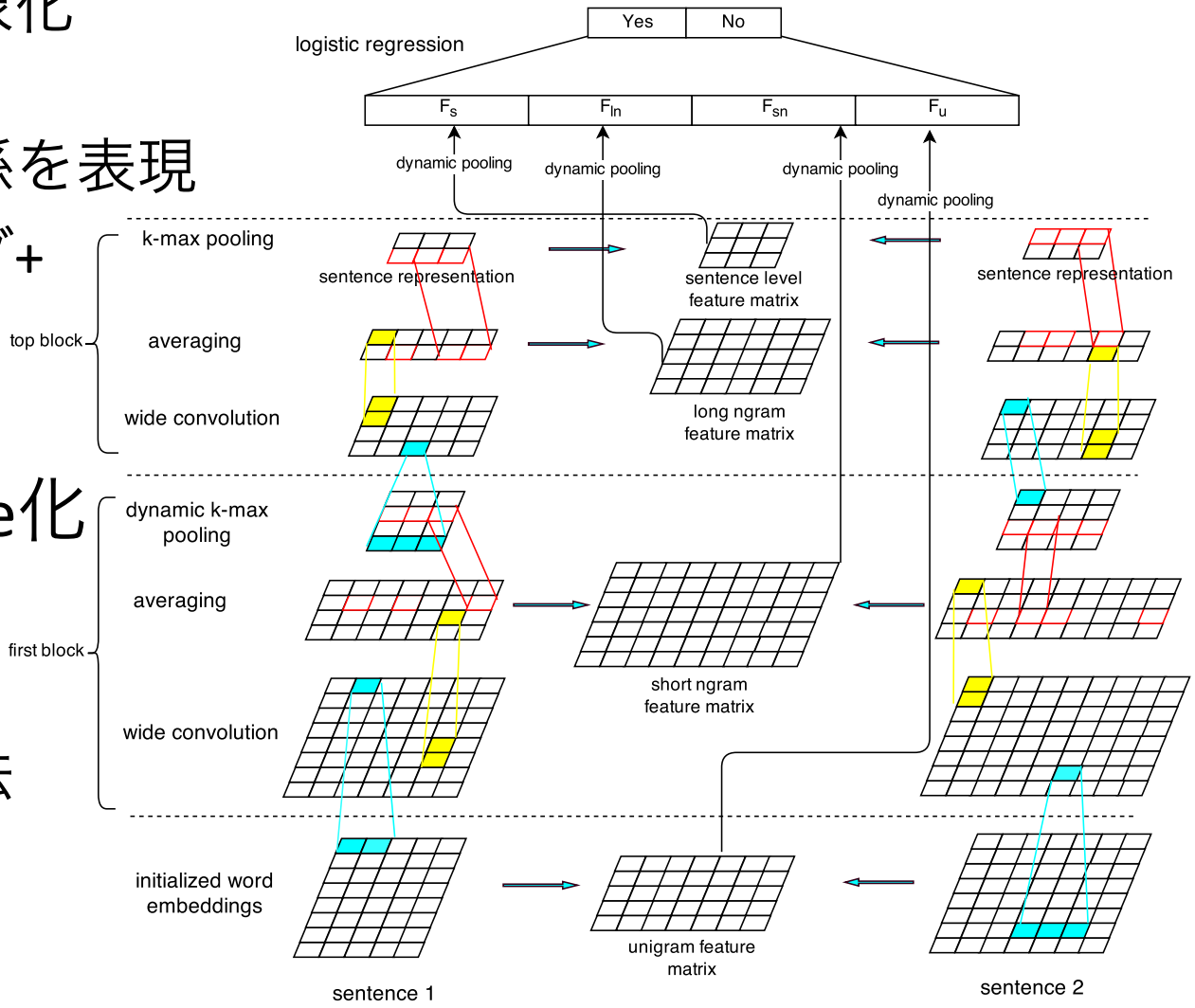
言い換え認識手法の例 (2/3)

- 表層的な類似度のスタッキング
  - SVMs [Finch+, 05]
  - 線形回帰 [Madnani+, 12]



言い換え認識手法の例 (3/3)

- 右図のCNNs [Yin+, 15]
  - 語彙情報の抽象化
  - 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



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]
    - 行列・テンソル分解 [Ji+, 13]
    - CNNs [Yin+, 15][He+, 15]
  - 統語構造やリッチな言語資源を用いても微々たる改善幅
- 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コーパス [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

ヤブレガサは日本では、本州、四国、九州に分布し、山地の林下の斜面などに生育する。

ヤブレガサは日本では、北海道、本州、四国、九州に分布し、山地の林下の斜面などに生育する。

ID20
正解率14%

太陽暦とは、地球が太陽の周りを回る周期を基にして作られた暦で、ユリウス暦は太陽暦の一種である。

ユリウス暦は、地球が太陽の周りを回る周期を基にして作られた暦で、太陽暦の一種である。

ID91
正解率33%

第4飛行隊は、かつて存在した南アフリカ空軍の飛行隊である。

第3飛行隊は、かつて存在した南アフリカ空軍の飛行隊である。

ID65
正解率48%

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%

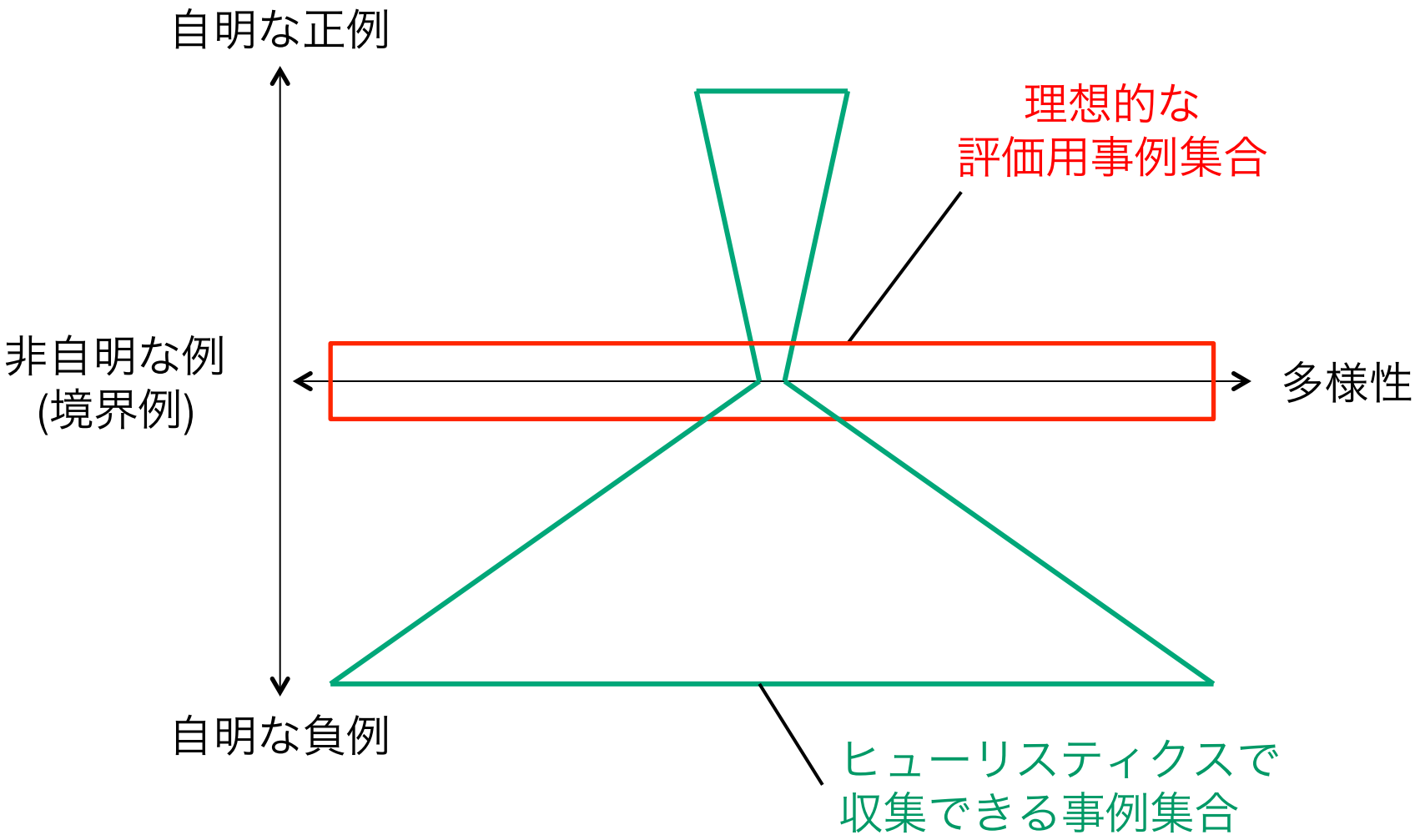
言い換え認識技術の評価用データの要件

- 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: 担保可能(と思われる)
    - 自明な事例だけでは意味がない
    - 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 [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

評価用データの作り方の比較

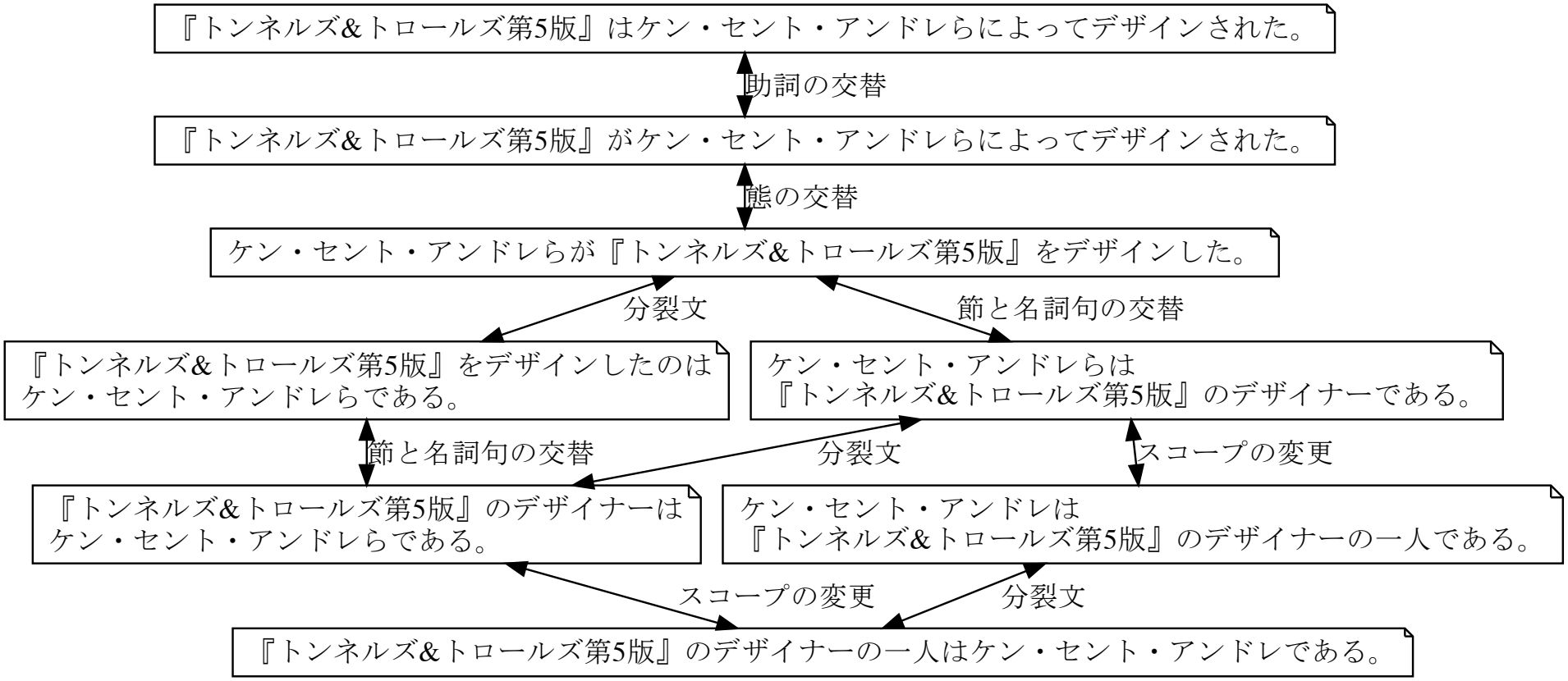
- ある範囲のテキストセットにアノテーション
  - 形態素解析，構文解析，固有表現抽出
  - 語義曖昧性解消，情報抽出
  - 述語項構造解析，照応解析，レビュー解析
  - 日本語校正，英語校正
- ある範囲の問題セットに模範解答を付与
  - 機械翻訳，自動要約 (1つないし複数)
  - 情報検索 (見つけられる限り)
  - 言語生成，言い換え生成 (無数)
- 言い換え認識ではあたりをつけてから人手で選別
  - 任意のテキスト対はほぼ間違いなく言い換えではない
  - 標本バイアスの回避 vs 人的コストの低減

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



もくじ

(5) pp.77-84 (8)

(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.33-44 (12)
(3) pp.46-56 (11)

(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，ベンチマーキング，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，データベース構築

(4) pp.58-75 (18)

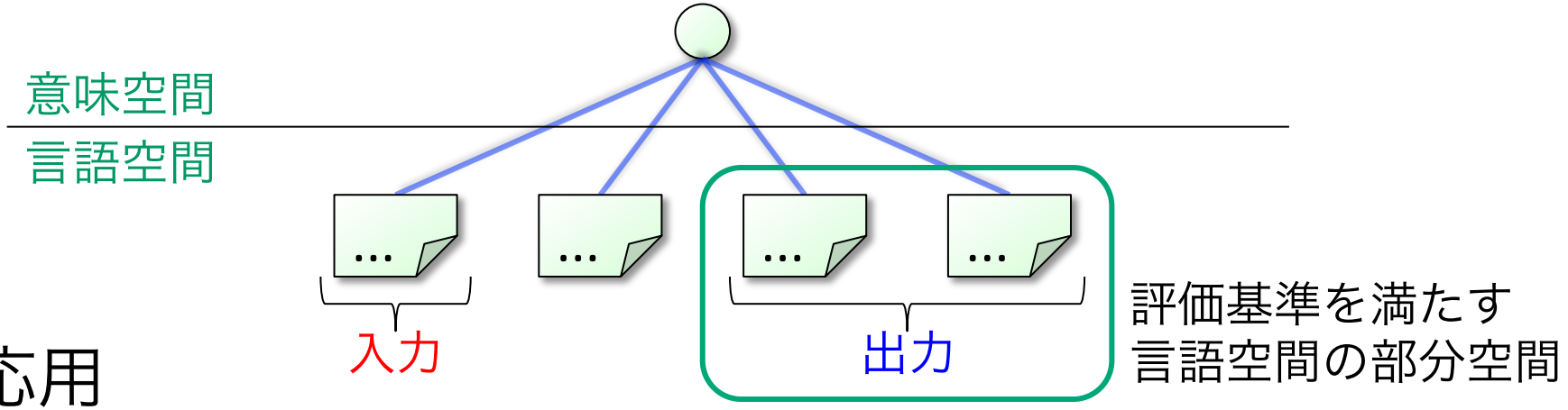
(L2) 事例研究
事例の収集と分析，計算アルゴリズム，必要な知識の同定，誤り分析

(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.16-31 (16)
(6) pp.86-105 (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
  - 簡単化，文圧縮，MTの前処理，TTSの前処理
- 評価方法
  - 出力を得るたびに人間が評価 (≒ MTの人間評価)
  - 少数の参照出力を事前に準備するアプローチは賛否両論

他のタスクとの関係

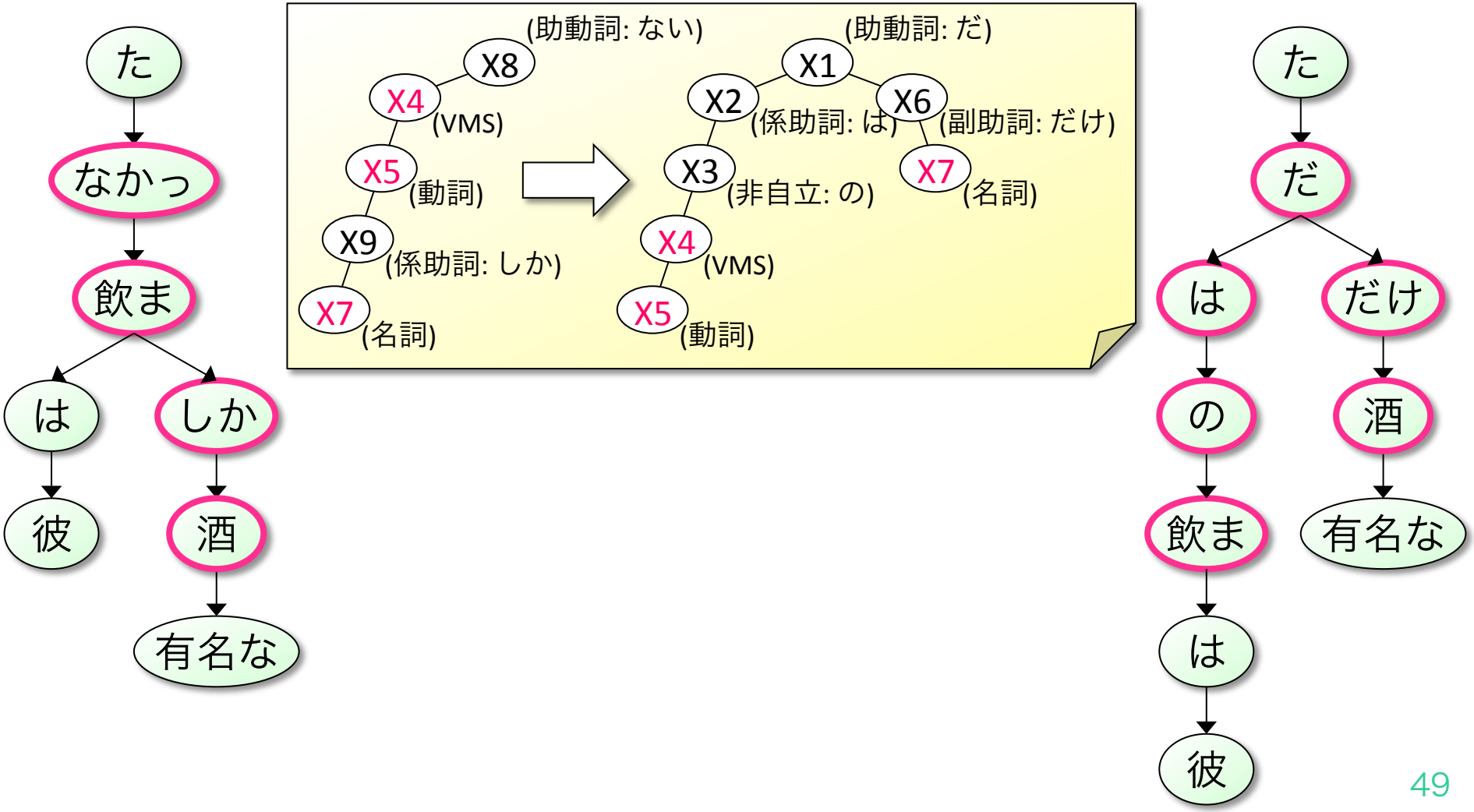
- text-to-text: MT, 文書要約, 簡単化
  - MTとの違い: 言い換え生成はMTの特殊版
    - 単一の言語内の変換
    - 与えられた評価基準を満たす必要がある
    - 入力全体を書き換える必要はない
  - 文書要約, 簡単化との違い: 大部分が重複
    - 有用な言い換えとそうでないもの
    - 言い換え以外にも有用な操作, 処理がある
  - 共通点: 妥当な解は複数存在, 事前に網羅することは困難
- 言い換え認識
  - 言い換え生成可能 → 言い換え認識可能
    - 逆は現実的に不可能

一般的な処理の枠組み

- 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: 語の難易度，構文パタンの難易度
    - SMTの前編集: フレーズのカバレージ(OOV数)
    - 文圧縮: 文全体
- 仮説生成
  - 言い換えが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 ≒ 単言語機械翻訳
- テキスト評価
  - 目的や制約に照らして最適な候補を選択

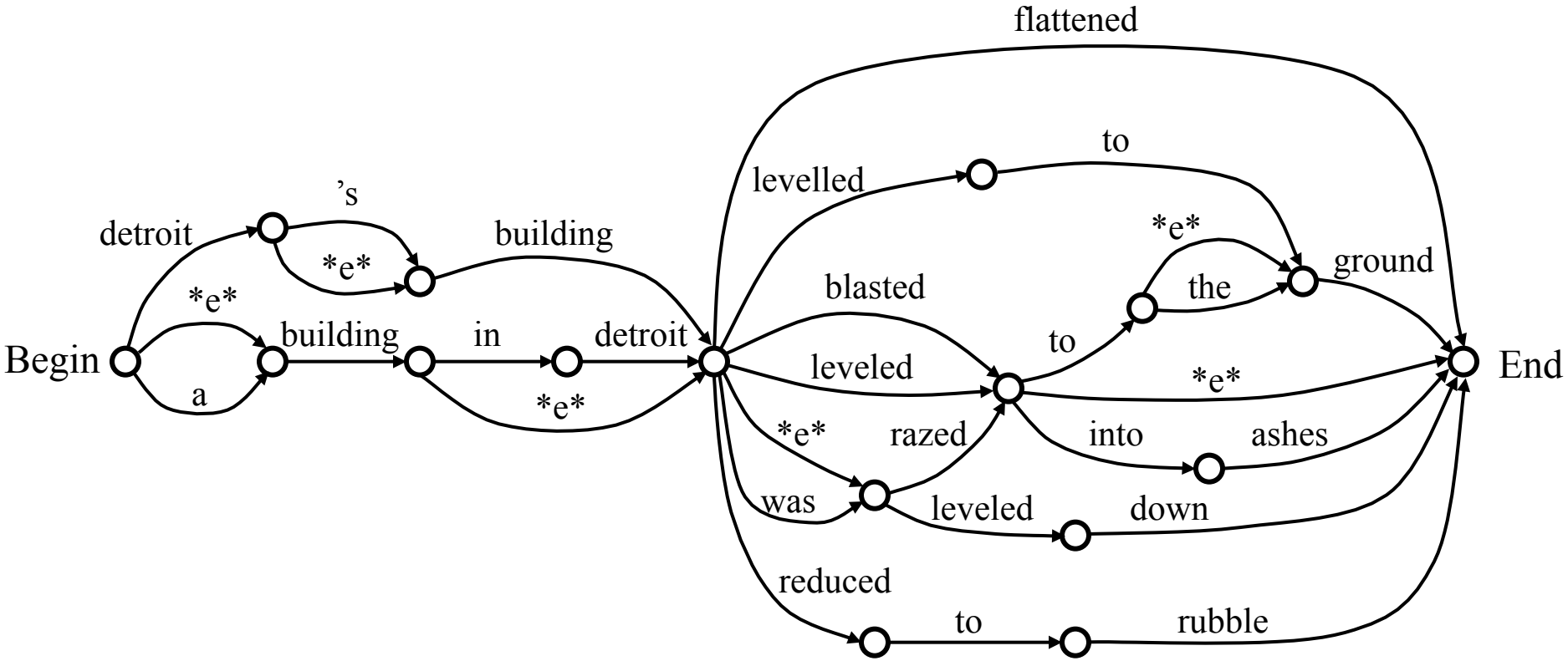
仮説生成手法の例 (1/3)

- 依存構造の構造変換 [Lavoie+, 00][Takahashi+, 01]
  - 必要な後処理の整理 [藤田+, 03]



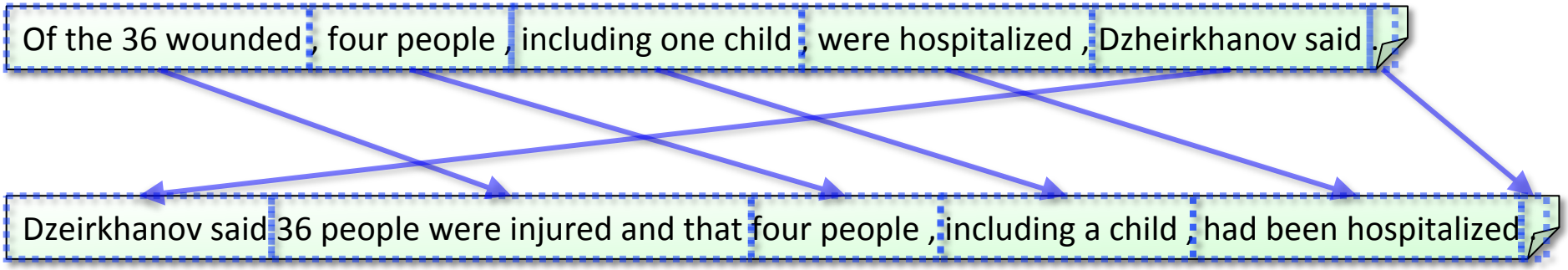
仮説生成手法の例 (2/3)

- Multiple-sequence alignment [Barzilay+, 03]
  - 単言語機械翻訳とみなすアプローチ
  - パラレル(コンパラブル)データが前提



仮説生成手法の例 (3/3)

- 統計的機械翻訳 [Quirk+, 04][Wubben+, 10]
  - 単言語機械翻訳とみなすアプローチ
  - パラレル(コンパラブル)データが前提



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
  - ✓ 事例研究にフィット
  - ✗ 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
  - ✓ フレームワークの数学的背景
  - ✓ 既存のツールを援用可能
  - ✗ 特定の一部分を書き換えるための制御が困難

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: 語の難易度，構文パタンの難易度
  - SMTの前編集: フレーズのカバレージ(OOV数)
  - 文書要約: 文長
- 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)
- 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. SMT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

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

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
    - 妥協案: MTと同様に少数の参照出力を用意する
  - 結局は目的依存か
    - 文章執筆支援: 候補を提示できればあとは人間任せ
    - SMTの前編集: 非文法的でもなんとかなる
- 仮説生成部の改良
  - 頑健性: 多様な言い換えのカバー
    - 必要な知識の整備
    - 言い換えの種類ごとに適した手法の高度化と統合
  - 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

言い換え生成もNN化?

- 言語生成/text-to-textにおけるNNモデル
  - MT [Kalchbrenner+, 13][Sutskever+, 14]
  - 文書要約(文選択+語選択) [Cheng+, 16]
  - 言語生成
    - 画像に対するキャプション生成 [Vinyals+, 15]
    - 対話生成 [Wen+, 15]
- 言い換え生成でも?
  - 前提: 正解データの存在
    - i.e., 唯一解を定めるという妥協
    - e.g., MSRPコーパス
  - 今年あたり注意型NNが出るかも
  - 仮説生成に適用できそうだが複数解を出力できるか?
  - 言い換え前後のテキスト評価も含められる?

もくじ

(5) pp.77-84 (8)

(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.33-44 (12)
(3) pp.46-56 (11)

(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，ベンチマーキング，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，データベース構築

(4) pp.58-75 (18)

(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.16-31 (16)
(6) pp.86-105 (20)

言い換えの処理に必要な知識 [藤田+, 04]

- すでに必要だと分かっているもの
  - 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)

      窮地 ↔ 苦境　　重傷 ↔ 大ケガ

      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，反義語辞書，特質構造 (生成語彙論)
- 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

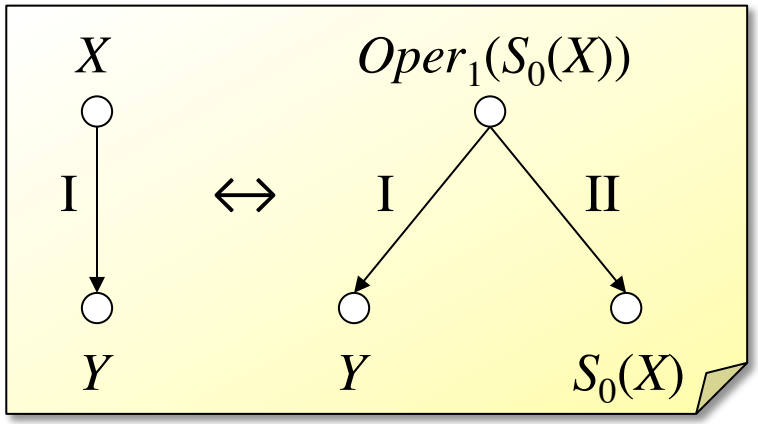
古典的な言い換え知識

- 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

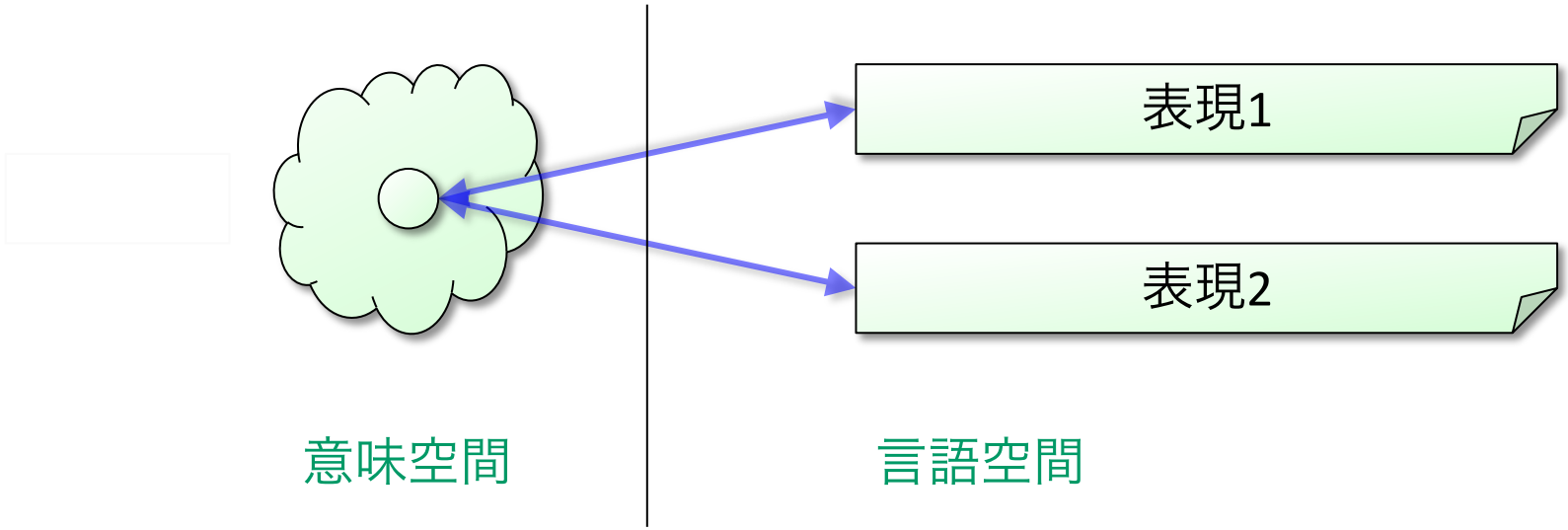
  - Meaning-text Theory [Mel'čuk+, 87]



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



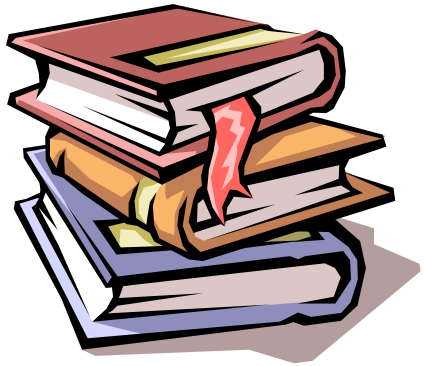
- 人間が設定した概念ラベル: Synset ID, 辞書の語釈文
- 文脈: bag-of-(周辺の表現), bag-of-(対訳フレーズ)
- 同一の情報源: 原文書, イベント, 動画
- etc.

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]

圧縮する: 気体や固体に圧力をかけてその体積を小さくする。

⇒ XがYを圧縮する ↔ XがYの体積を小さくする

- 同じ語に対する複数の辞書の語釈文 [村田+, 04]

あべこべ

順序・位置などの関係がさかさまに入れかわっていること。

順序・位置・関係がひっくり返っていること。

- 最大の課題
  - 人間向けの辞書が整備されているという前提

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

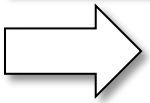
上位・下位 liquor ↔ wine

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

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



thrown into jail ↔ imprisoned

under control ↔ in check

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
    - 小説 [Barzilay+, 01]
    - MTの自動評価用の複数参照訳 [Pang+, 03]

At least 12 people lost their lives in last week's fighting.

Last week's fight took at least 12 lives.

The fighting last week killed at least 12.

- 単言語コンパラブルコーパス
  - 動画内容の言語化 [Chen+, 11]



A person breads a pork chop.

A woman is coating a piece of pork with breadcrumbs.

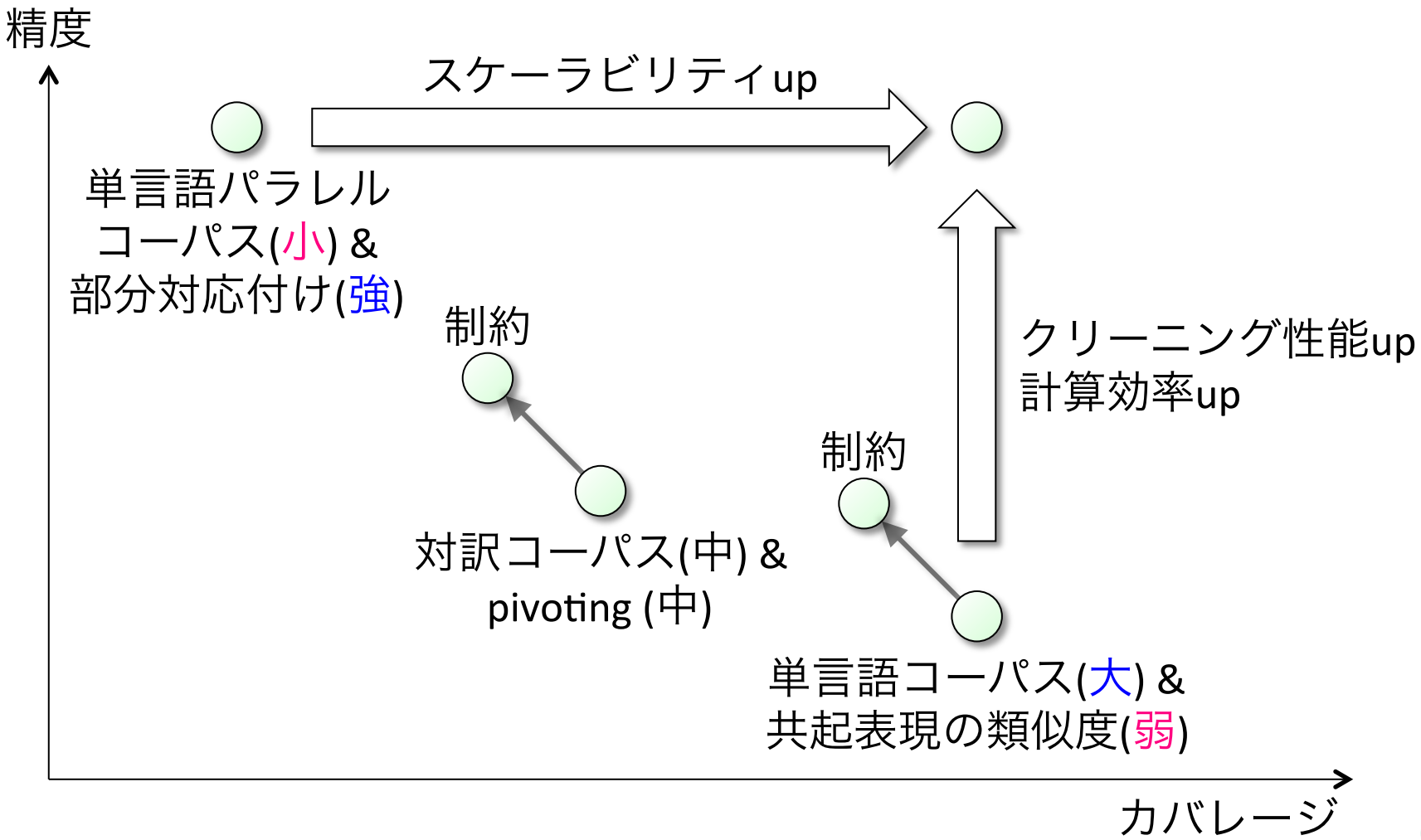
同一のイベントについて述べた複数の記事

- 単言語コンパラブルコーパスの自動構築
  - 異なる通信社の記事がたくさんあり，ずっと増え続ける
    - 同じ内容である保証はないが類似する情報を含む
  - Google News の見出しの対応付け [Wubben+, 09]
  - 編集距離に基づくクリーニング [Dolan+, 04]

- 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]

言い換え知識獲得の課題

- カバレージと精度の両立
  - 制約の導入は上澄み抽出に過ぎない



スケールしそうなアプローチ (1/2)

- 単言語コーパス → コンパラブルコーパス化
  - 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]
    - 言い換えDBと言っておきながら…



語彙的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

語彙的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%

利用可能なリソース (1/2)

- 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: 語彙的に拡張されたPPDB [Fujita, 16]
    - 英語: 6.9億表現対
    - 仏語: 9.4億表現対
    - 西語: 6.3億表現対
    - 独語: 5.3億表現対

利用可能なリソース (2/2)

- 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]
- 注釈付きコーパス
  - MSRP: MSR 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]
  - PIT: Twitter 言い換えコーパス (英語) [Xu+, 14]
  - NTCIR RITE-2: 日本語含意関係認識データ [Watanabe+, 13]

もくじ

(5) pp.77-84 (8)

(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.33-44 (12)
(3) pp.46-56 (11)

(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，ベンチマーキング，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，データベース構築

(4) pp.58-75 (18)

(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.16-31 (16)
(6) pp.86-105 (20)

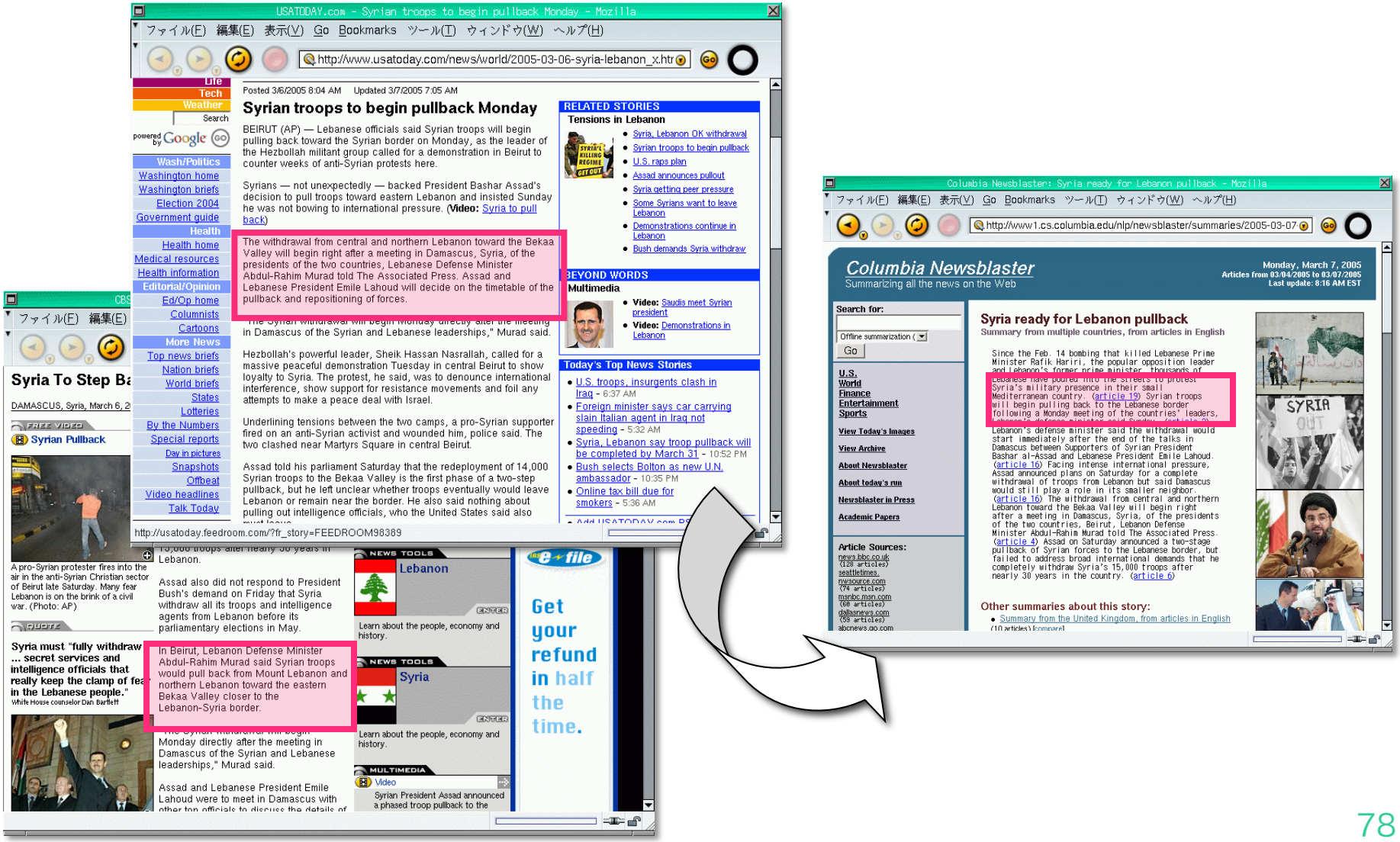
応用技術と言い換え

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

文書要約

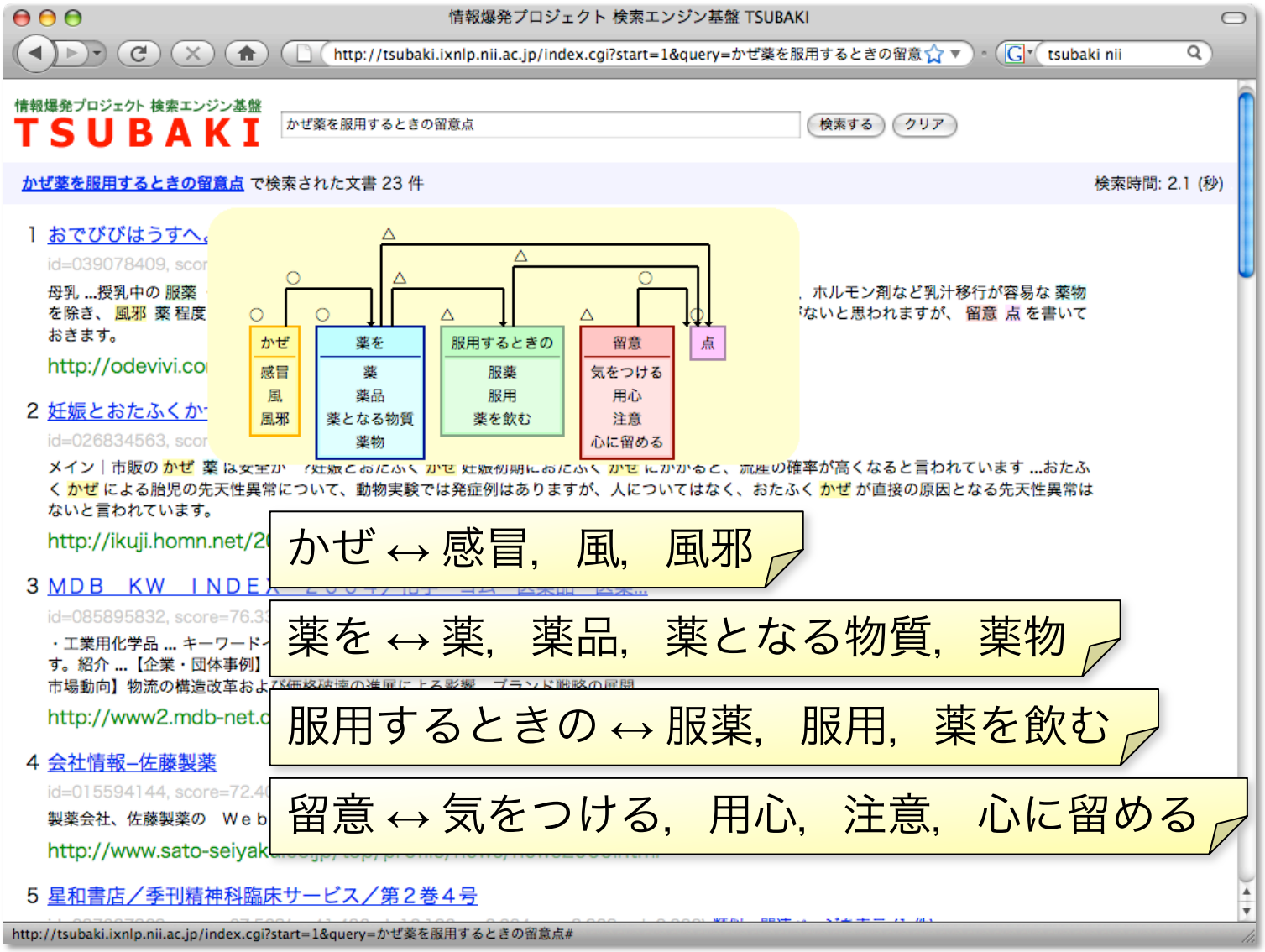
人間向け言い換え認識・生成

- e.g., 複数文書要約 [Barzilay, 03]



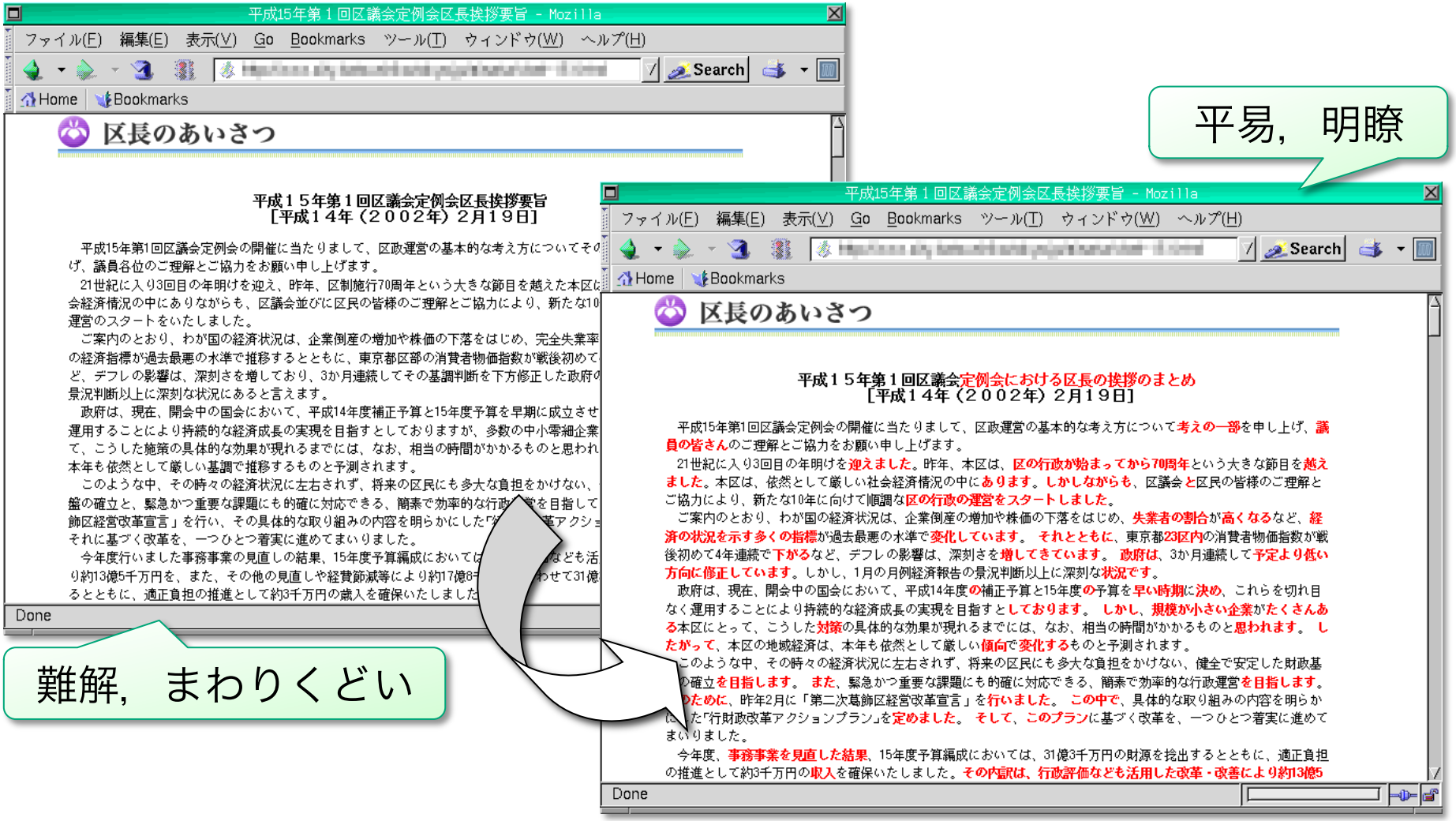
人間向け言い換え認識

- e.g., 情報検索 [Shibata+, 08]



人間向け言い換え生成

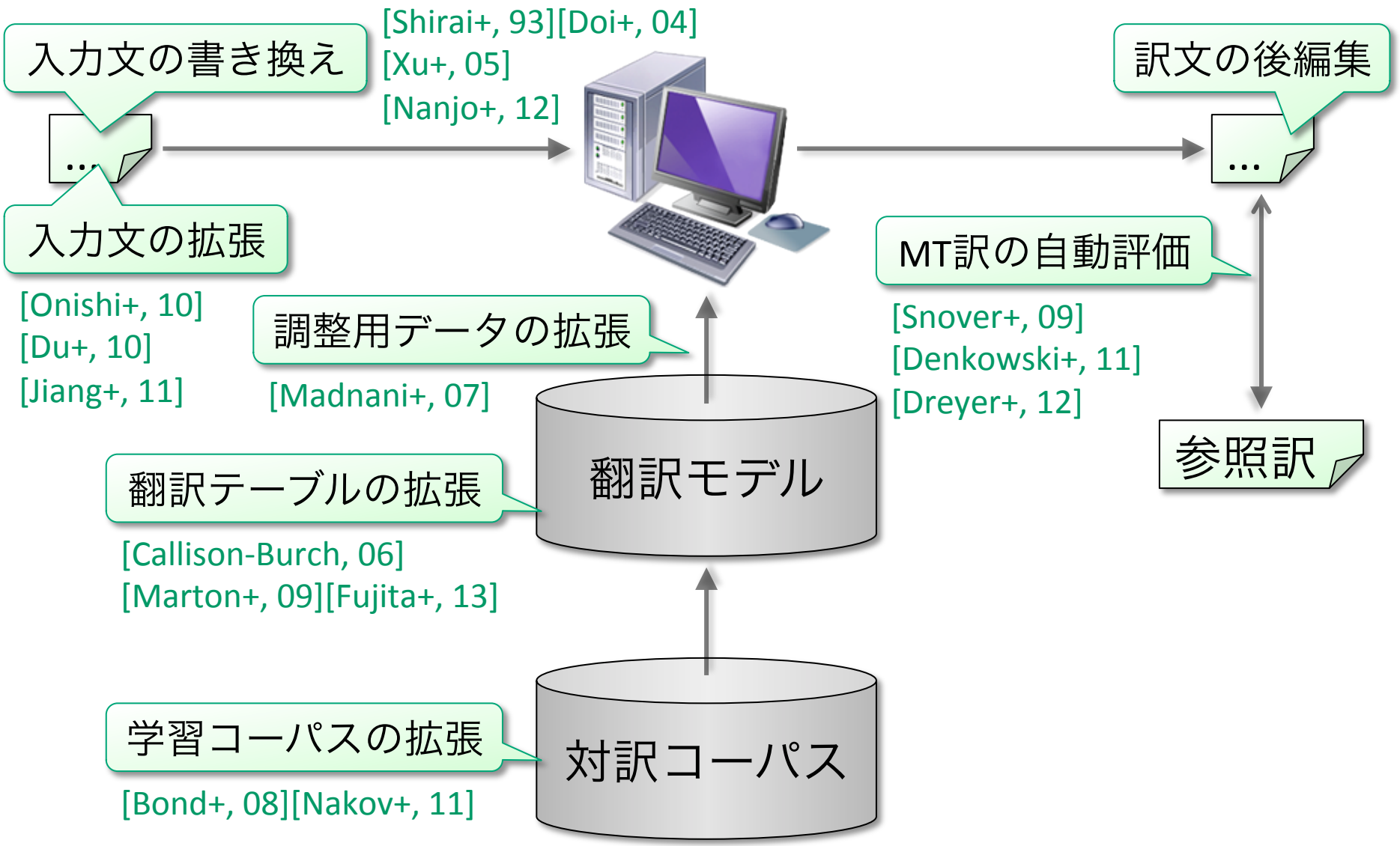
- e.g., ろうあ者向けのテキスト簡単化 [Inui+, 03]



テキスト簡単化 ≠ 言い換え

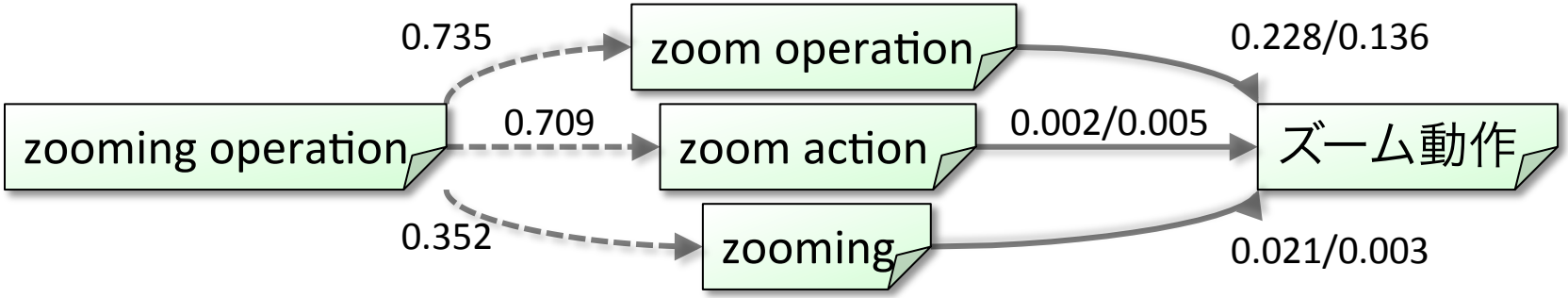
- 言い換え (のサブセット) である必要はない
  - 伝達内容の省略
    - 言語表現の難解さは制御できても
    - 難解な概念はどうしても伝えられない → 抽象化，省略
  - 情報の構造化・可視化
    - e.g., 箇条書き，文字装飾，図示
- 用途に合わせた簡単化
  - 制限言語 [Huijsen, 98]: 翻訳しやすい
  - 産業日本語 [熊野+, 09]: 人間・計算器の解釈が容易
  - やさしい日本語 [庵+, 13]: 日本語非母語話者でも分かる
    - e.g., NHK “News Web Easy” [田中+, 13]
  - 小学生向け語彙平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

機械翻訳研究への各種応用

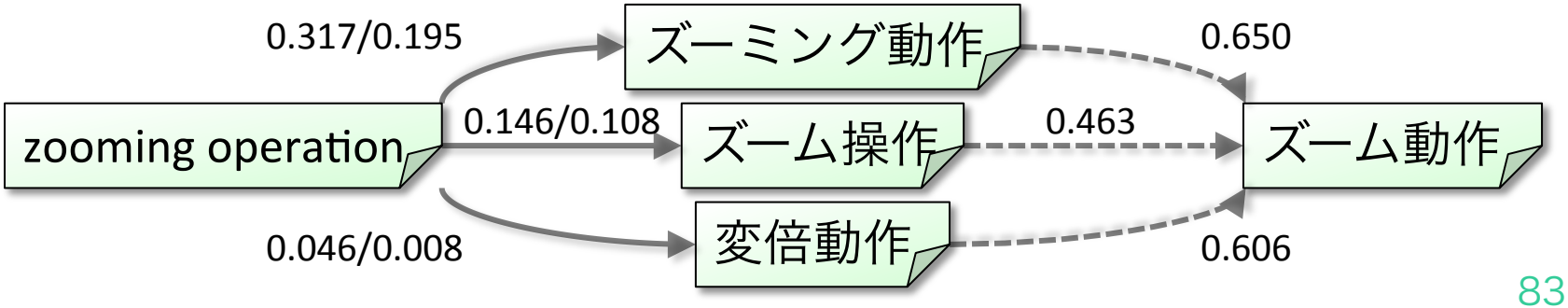


翻訳テーブルの拡張

- 対訳コーパスの限界
  - 対訳コーパスに出現した表現しか翻訳できない
  - 対訳コーパスに出現した表現にしか翻訳できない
- 言い換えと翻訳テーブルの組み合わせ
  - 未知の表現に対する翻訳の生成 [Callison-burch+, 06]



  - より多様な目的言語表現の生成 [Fujita+, 13]



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.77-84 (8)

(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.33-44 (12)
(3) pp.46-56 (11)

(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，ベンチマーキング，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，データベース構築

(4) pp.58-75 (18)

(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

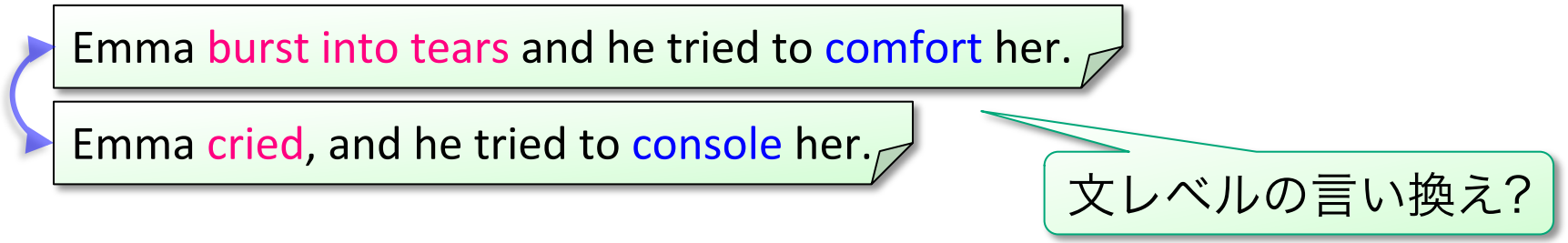
(1) pp.16-31 (16)
(6) pp.86-105 (20)

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]

言い換えの分類における課題

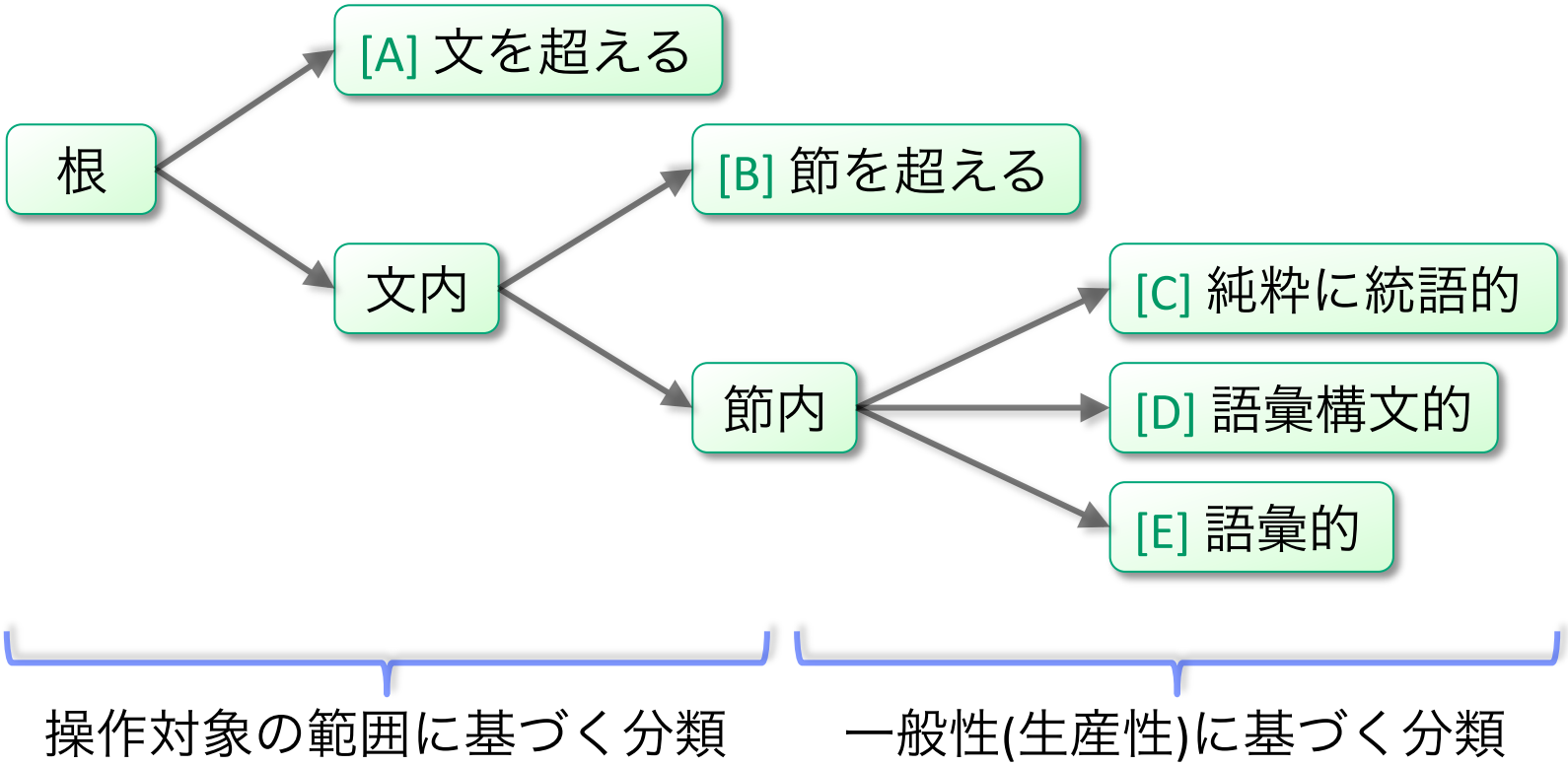
- 原子性



- 網羅性
  - 自動獲得結果の分析のみではなく，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

- 5種類の大分類と60種類の小分類 [Fujita, 10]
  - cf. 乾解説論文 [乾+, 04]
  - cf. 『言い換えのあれこれ』 http://paraphrasing.org/



[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。

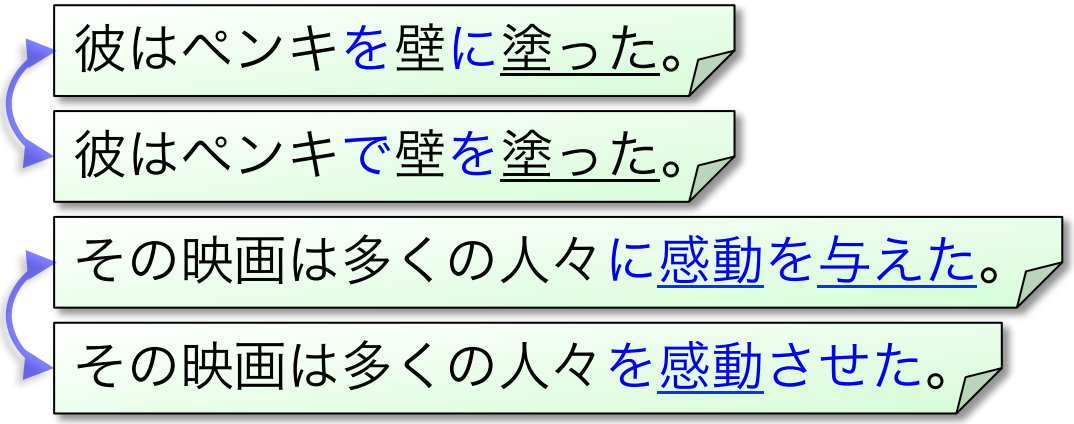
彼が飲んだのは有名な酒だけだった。

[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

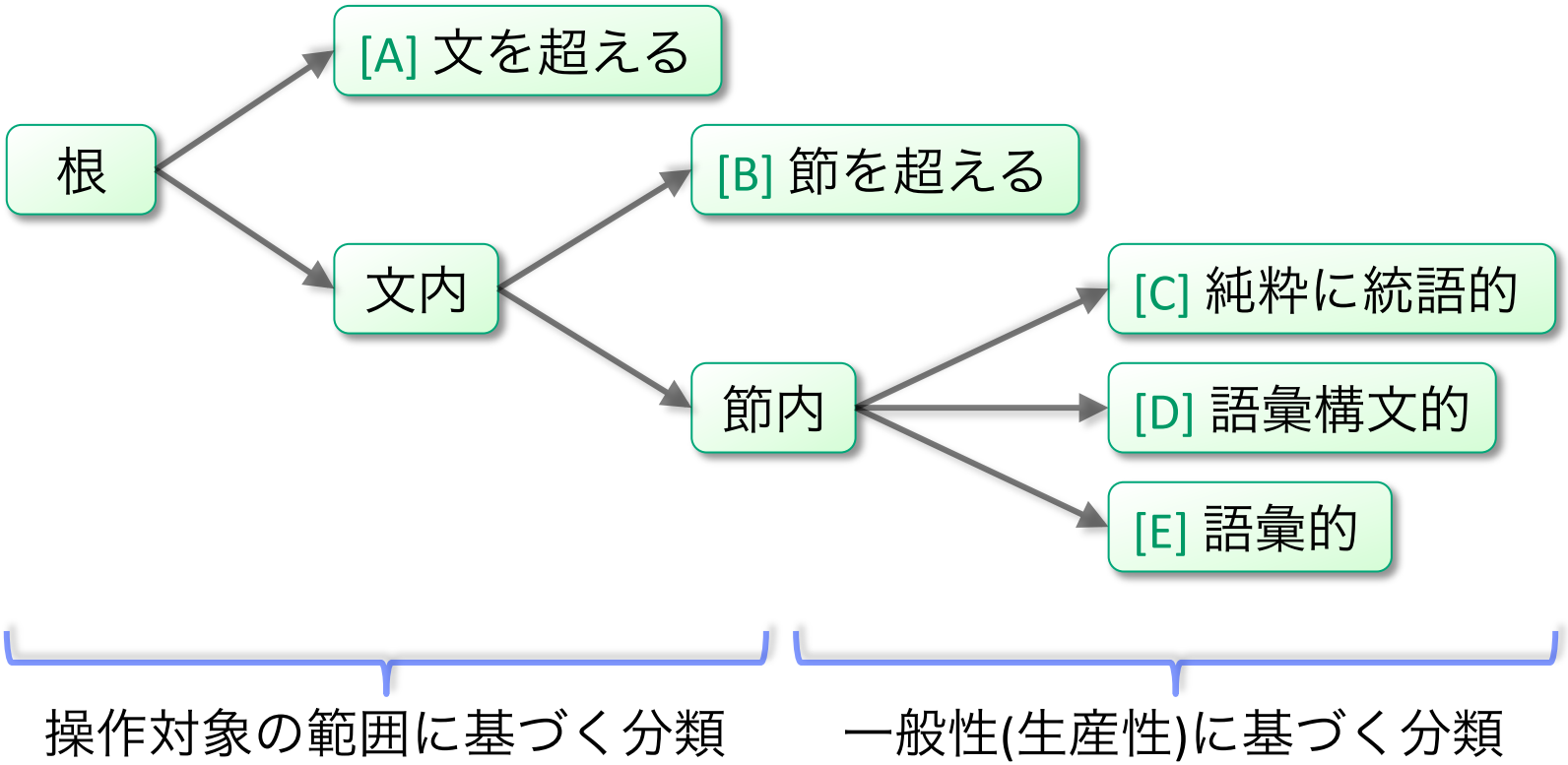
- 機能動詞構文 (名詞 ↔ 動詞)
  - その映画は多くの人々に感動を与えた。
  - その映画は多くの人々を感動させた。
- 形容詞 ↔ 動詞
  - 風が強くなってきたようだ。
  - 風が強まってきたようだ。
- 名詞 ↔ 形容詞
  - 日焼けのため肌が赤みを帯びている。
  - 日焼けのため肌が赤くなっている。

[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
お過ごしください。
  - 週末を有意義に
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
  - 語彙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試合振りに勝点を得た。
    - 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。

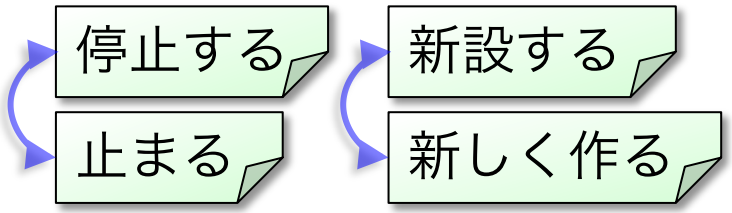
議論 (2/3): 各々の実現方法

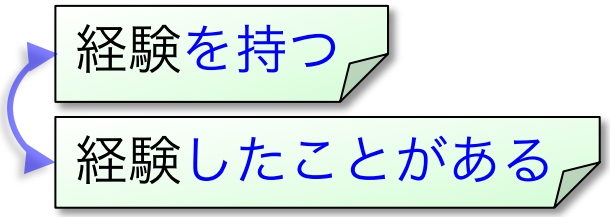
- [A-C] 種類は少ない → 過剰生成を許容しつつ書く
- [D] 統語と意味の相互作用を要解明 → 語彙意味論
  - Meaning-Text Theory [Mel'čuk+, 87]
  - 語彙概念構造 [Jackendoff, 91]
  - 生成語彙論 [Pustejovsky, 95]
  - FrameNet [Baker+, 98]
  - Verb classification [Levin, 93]
    → VerbNet [Kipper+, 00]
- [E] 要丸覚え → 自動獲得に頼るしかない
  - 既存の語彙資源も最大限に活用
    - e.g., 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

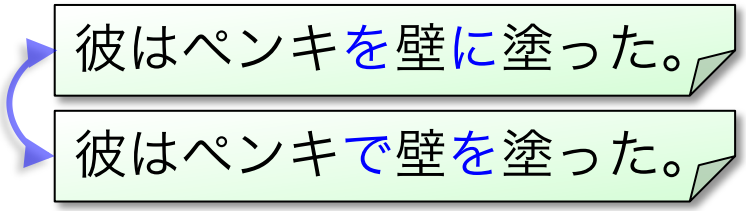
大規模な言い換え知識獲得によって，[E]と同じように扱える見込みが出てきた．

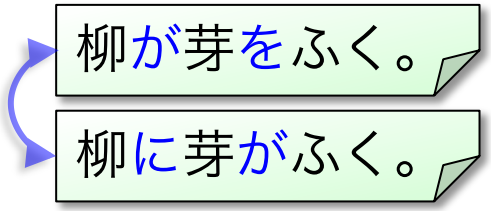
議論 (3/3): 現象を網羅することの難しさ

- 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



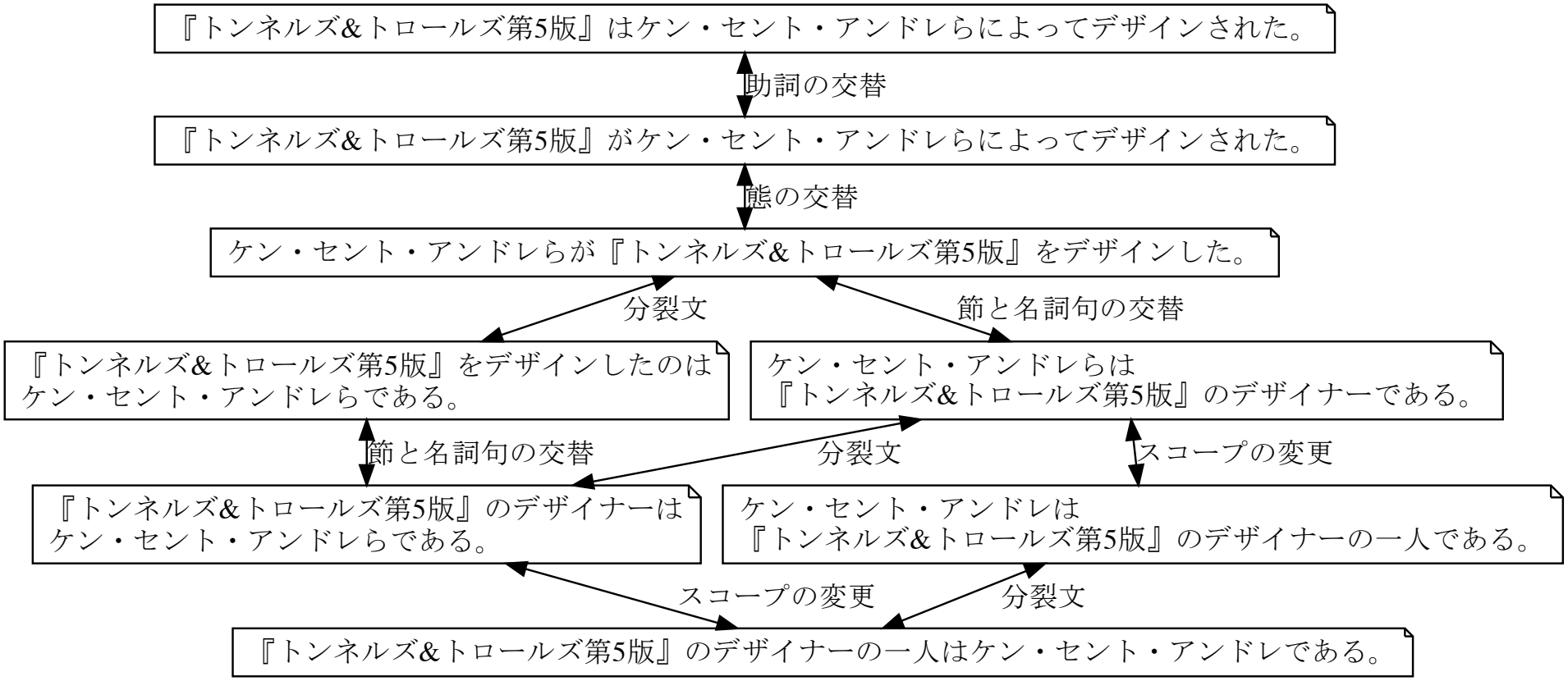






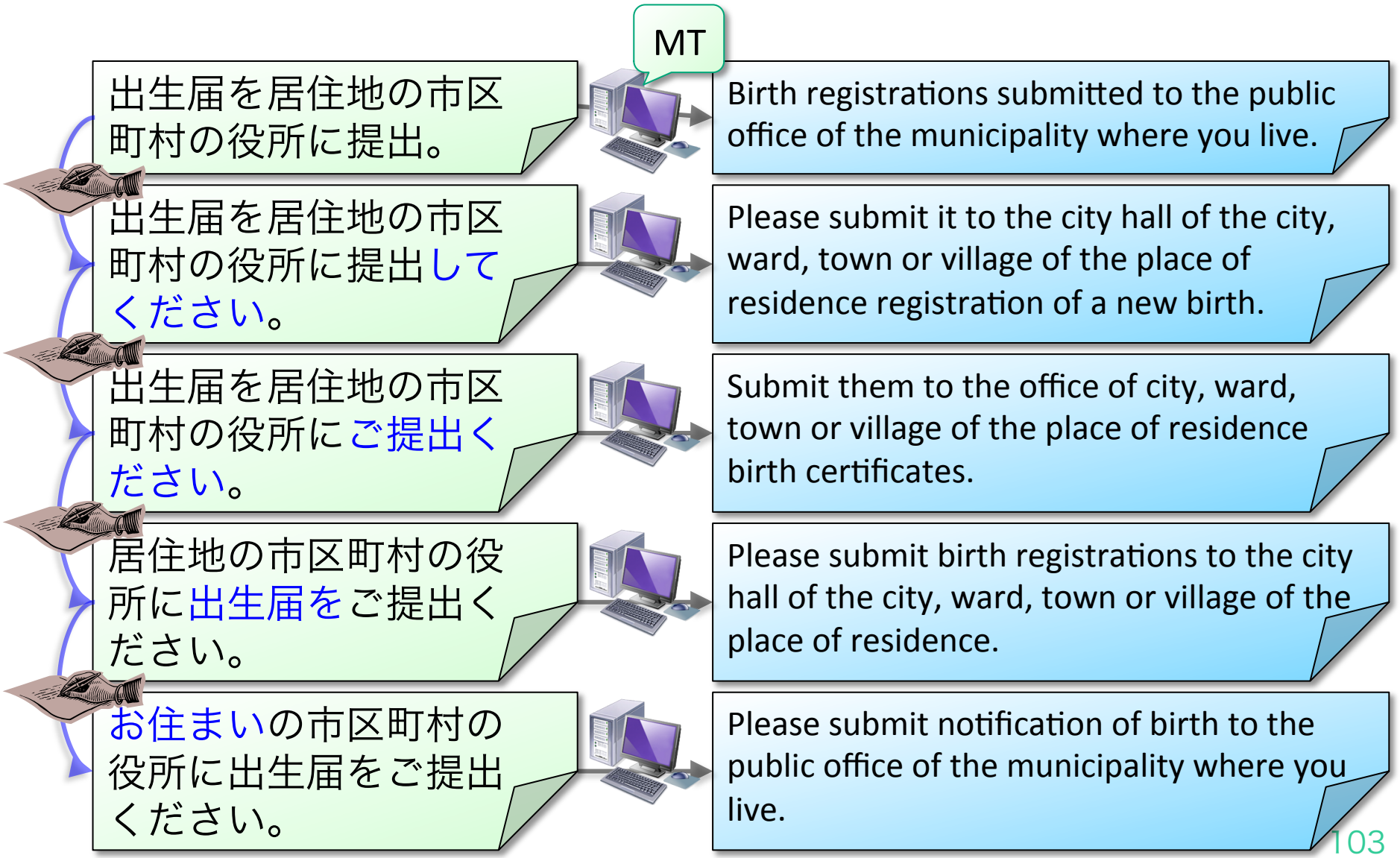
我々の試み (1/2)

- 既存のデータ中の事例の分解・分類 [藤田+, 15]



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]



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

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。

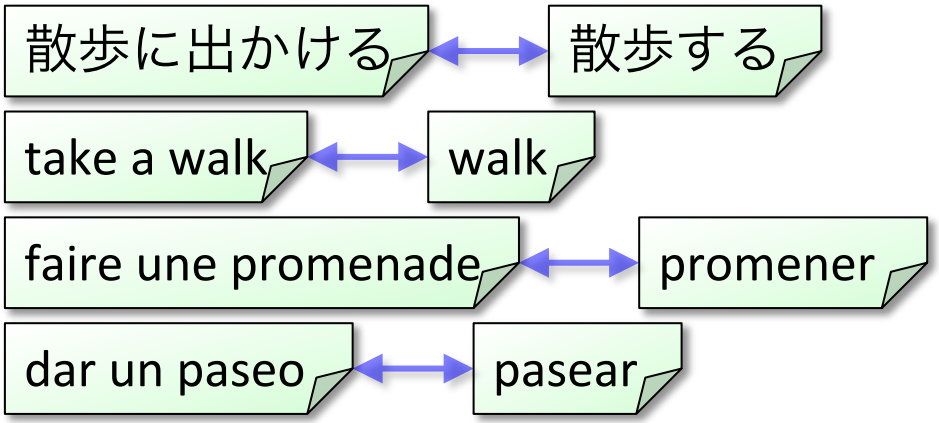
残っている研究課題
(まとめに代えて)

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



Further Reading

[Androutsopoulos+, 10] Ion Androutsopoulos and Prodromos Malakasiotis. A Survey of Paraphrasing and Textual Entailment Methods. Journal of Artificial Intelligence Research, Vol. 38, pp. 135-187, 2010.

[Bhagat+, 13] Rahul Bhagat and Eduard Hovy. What is a paraphrase? Computational Linguistics, Vol. 39, No. 3, pp. 463-472, 2013.

[藤田+, 15] 藤田篤, 柴田知秀, 松吉俊, 渡邉陽太郎, 梶原智之. 言い換え認識技術の評価に適した言い換えコーパスの構築指針. 言語処理学会第21回年次大会ワークショップ『自然言語処理におけるエラー分析』発表論文集, ID:20, 11 pages, 2015.

[乾+, 04] 乾健太郎, 藤田篤. 言い換え技術に関する研究動向. 自然言語処理. Vol. 11, No. 4, pp. 151-198, 2004.

[Madnani+, 10] Nitin Madnani and Bonnie Dorr. Generating Phrasal & Sentential Paraphrases: A Survey of Data-Driven Methods. Computational Linguistics, Vol. 36, No. 3, pp. 341-387, 2010.

[Paraphrasing.org, 13] Atsushi Fujita. 言い換え技術に関する研究動向の包括的調査. http://paraphrasing.org/

[Recasens+, 10] Marta Recasens and Marta Vila. On Paraphrase and Coreference. Computational Linguistics, Vol. 36, No. 4, pp. 639-647, 2010.

[Vila+, 11] Marta Vila, M. Antònia Martí, and Horacio Rodríguez. Paraphrase Concept and Typology: A Linguistically Based and Computationally Oriented Approach. Procesamiento del Lenguaje Natural, No. 46, pp. 83-90, 2011.