

言い換え技術の研究動向: 分類体系, 知識獲得, 応用

Atsushi Fujita



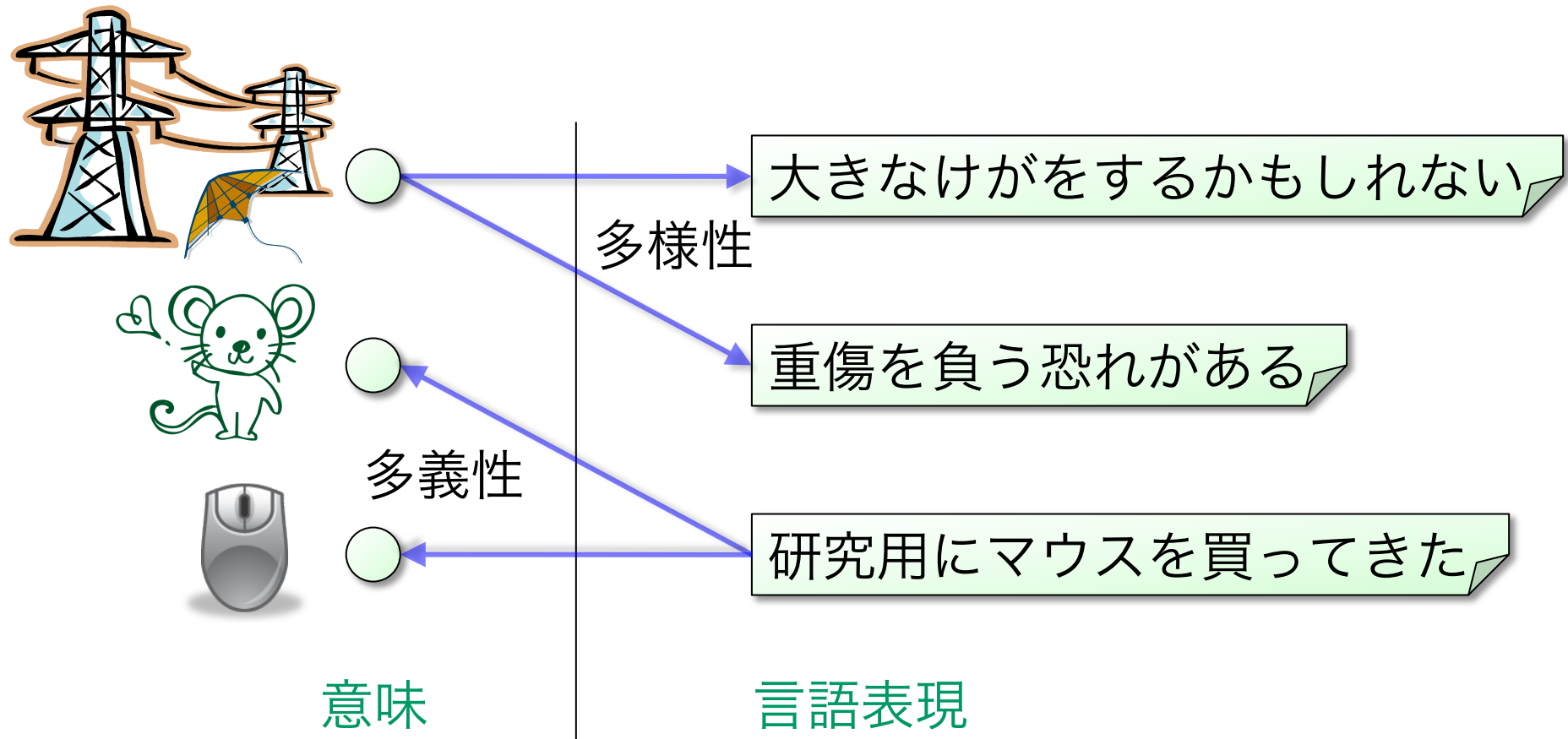
公立はこだて未来大学

<http://paraphrasing.org/~fujita/>

今日の話題: 言い換え

■ 換言, 言い替え, Paraphrase (Paraphrasing)

- 同じ意味内容を表す, 同言語の異なる言語表現
- ある表現の言い換えを生成する行為

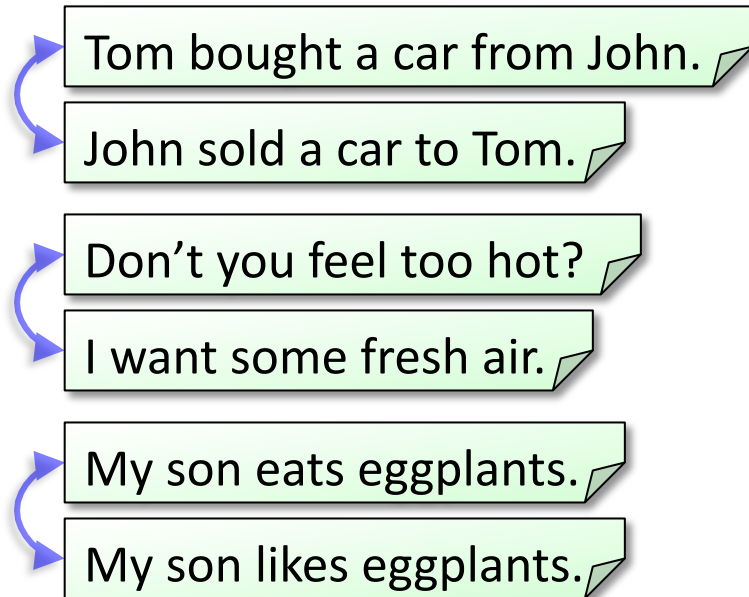


本講演の趣旨

- 前稿解説論文 [乾+, 04] の内容の再吟味
- それ以降の研究の整理
 1. 分類体系:
同義性のレベル, 関連する概念, 分類の観点
 2. 言い換え知識の獲得・蓄積:
規模, 精度, 獲得可能な種類による各手法の特徴付け
 3. 応用:
生成か認識か, 人間向けか機械処理向けか
- キーメッセージ
 - データドリブンが隆盛も, まだまだ道半ば
 - 業界全体で大事なことを見落としている可能性!?
 - ご意見募集, 人材募集

今日の話題1. 言い換えの分類体系

■ これは言い換え？



■ 分類の意義

- 人間が行っている計算の種類の俯瞰
- 自動化のための基盤
 - 必要なリソースの検討
 - 目的に応じて選択的に使えるようなモジュール化

今日の話題2. 言い換え知識の獲得・蓄積

■ 頑健な処理には大規模な知識ベースが不可欠

- 人手でコツコツ作成
- 自動獲得

■ 各種言い換え知識獲得手法の俯瞰

- 単言語コーパス + 分布仮説
 - 大きい → (相対的には) カバレッジが高い
 - 根拠の信頼性に乏しい → 精度が低い
- n言語パラレルコーパス + 翻訳テーブル
 - 文レベルの同義性が保証されている → 精度が高い
 - データは小さい → カバレッジが低い

今日の話題3. 応用

	言い換え生成	言い換え認識
人間向け	文章執筆支援	剽窃検出
	文章読解支援	翻訳メモリ検索
	TTS前処理	テキスト要約
	MT後編集	QA IR
	ステガノグラフィ	
機械処理向け		SMTの拡張 東ロボ!?
	MT前編集	IE DM

もくじ

- 
- 1. 意味の同一性
 - 2. 言い換えの分類体系
 - 3. 計算～知識獲得
 - 4. 言い換え技術の応用例
 - 5. 今後の研究課題

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

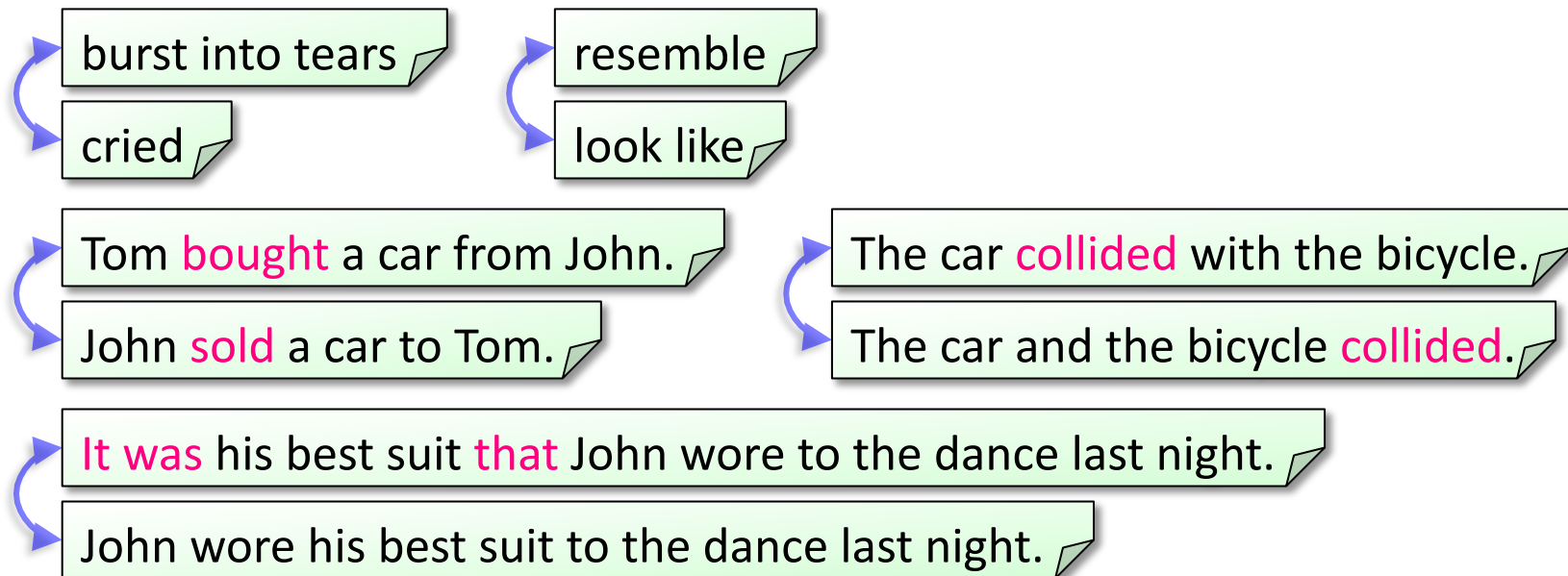
paraphrase _(v): to express in a shorter, clearer, or
different way what someone has said or written

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

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?



- 全貌は明らかにされていないし, 境界も不明瞭



言語表現の意味

■ 各種意味論 (semantics)

- 形式意味論
- 状況意味論
 - 談話表示理論 (DRT) [Kamp, 81] ... ロバ文
 - メンタルスペース理論 [Fauconnier, 85] ... アキレスと亀
- 語彙意味論
 - フレーム意味論 [Fillmore, 76][Baker+, 98] ... 述語項構造
 - 語彙概念構造 [Jackendoff, 90] ... 統語と意味のインタフェース
 - 生成語彙論 [Pustejovsky, 95] ... 特質構造, 換喩, etc.

意味処理の例題としての言い換え

■ 同義であるとは何か

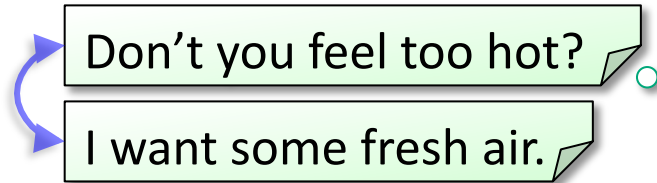
- “Paraphrases are **approximate** conceptual equivalence among outwardly different material” [De Beaugrande+, 81]
- “Every two forms (in language) **contrast** in meaning” [Clark, 92]

■ 同義性を考える3つのレベル [Sato, 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



新鮮な空気が欲しい、
だから窓を開けてくれ、
という依頼・指示

■ 言語行為論 [Austin, 62]

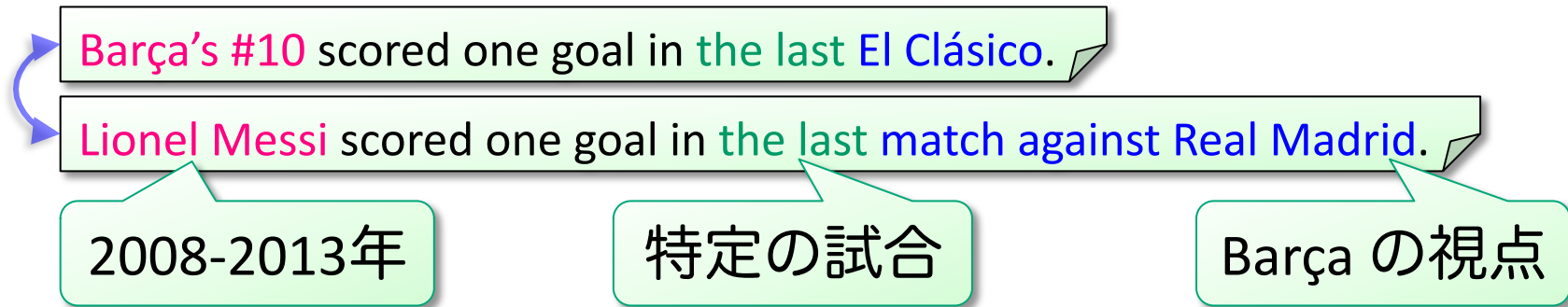
- 発語内行為 (Illocutionary act): 言葉の背後にある効力
- 発語媒介行為 (perlocutionary act): 聞き手への効果

■ 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

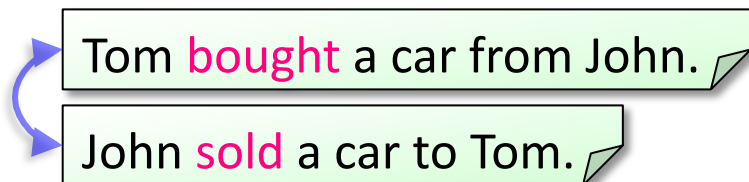
■ 共参照



- 異なる状況下では偽
 - e.g., **Ronaldinho**, **Riquelme**, **Rivaldo**, ...
 - e.g., **match against Barça**, **match between Barça and Real**
- 談話あるいはそれを超えた意味
 - 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あるいは命題論理



- 言語知識と世界知識に基づいて説明可能
 - 情報伝達の状況を参照することなく
- 言外の意味は異なりうる [Edmonds, 99][Inkpen+, 06]
 - 主題・焦点
 - 様式 (e.g., 形式性, 丁寧さ)
 - 感情・態度

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

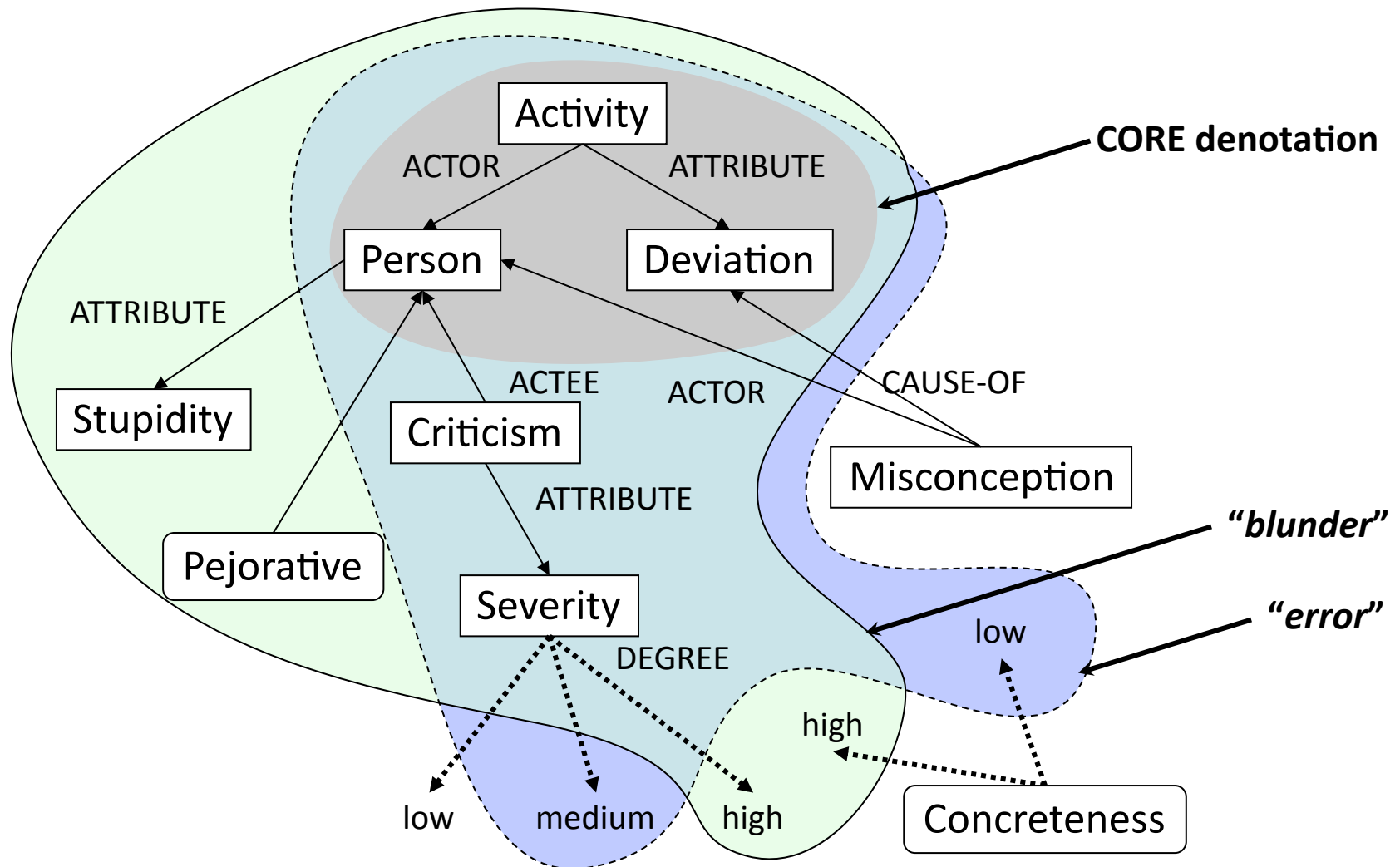
paraphrase _(v): to express in a **shorter, clearer**, or different way what someone has said or written

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 equivalence [De Beaugrande+, 81]
 - Quasi-paraphrases [Bhagat+, to appear]

“blunder” と “error” の比較

[Edmonds, 99]



違いを生じさせるねらい

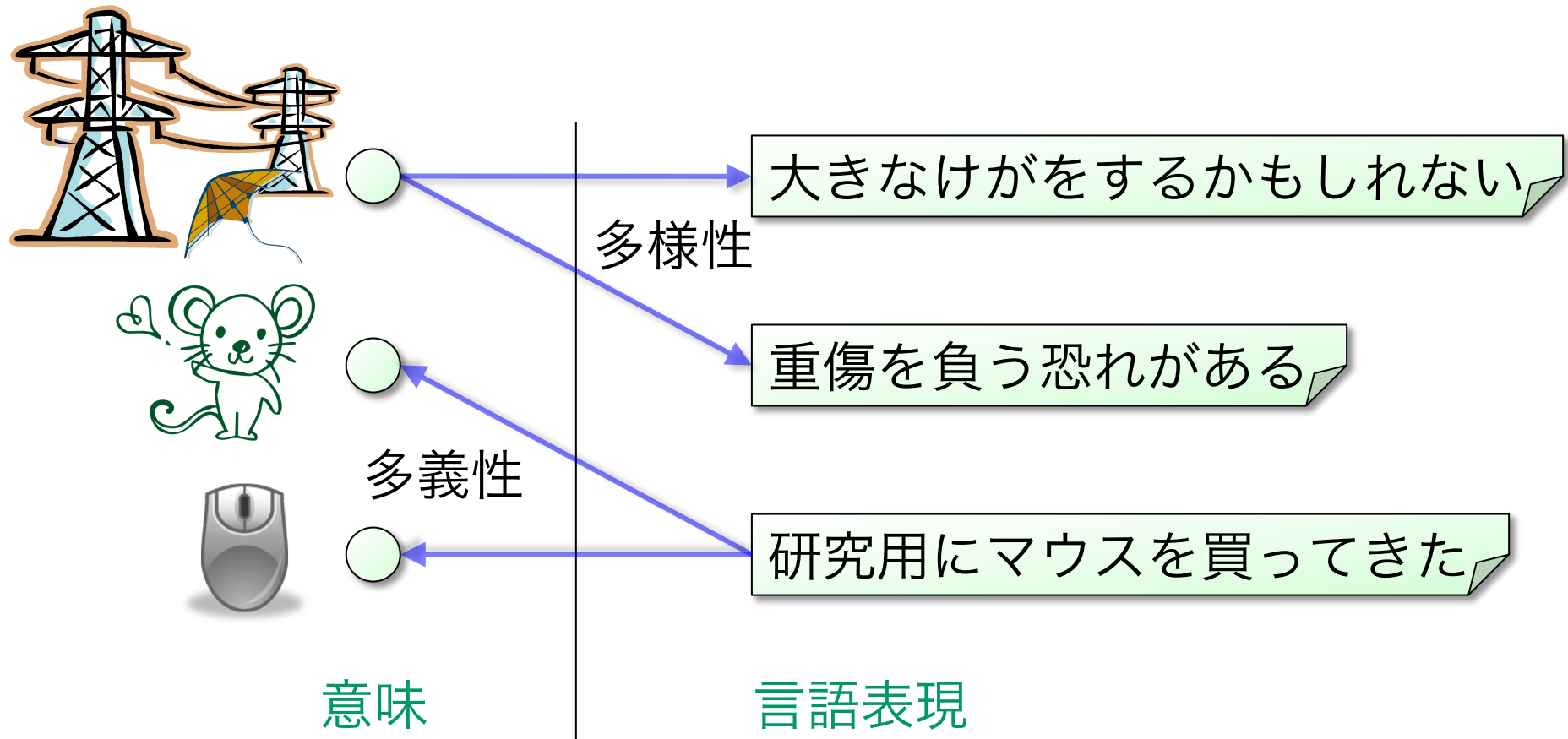
■ 発信者と受信者の間の情報伝達の促進

- 人間どうしの対話
 - 発話内容の確認 [高塚, 99]
 - 仲間うちにおけるメンツを保つため [国広, 00]
- 応用技術における目的
 - 簡単化: 難解 → 平易
 - 推敲: まわりくどい, 冗長 → 明瞭 (場合によっては逆)
 - 客観化: 感情的 → 非感情的
 - テキスト圧縮, 要約: 長い → 短い

今日の話題: 言い換え

■ 換言, 言い替え, Paraphrase (Paraphrasing)

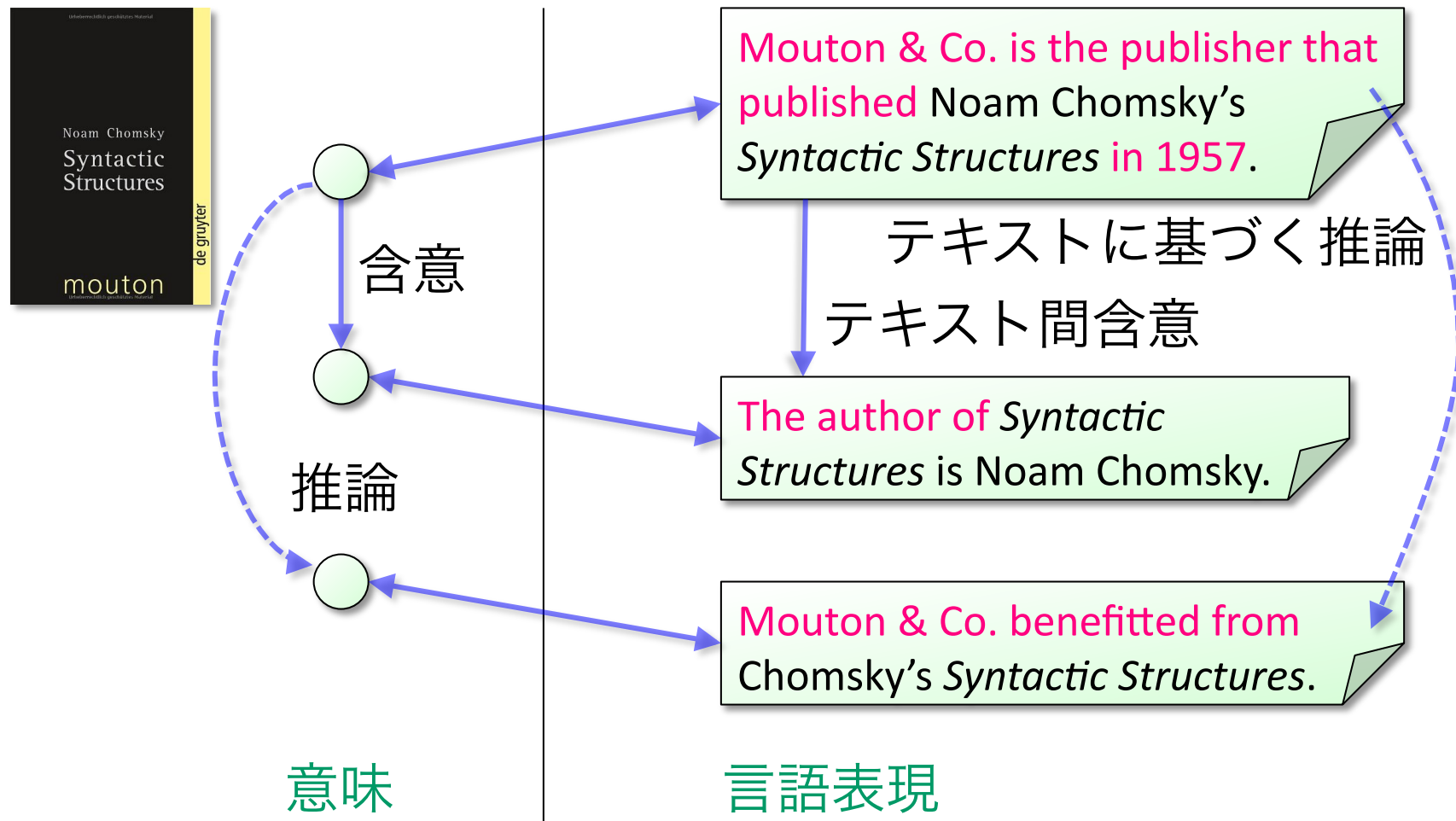
- 同じ意味内容を表す, 同言語の異なる言語表現
- ある表現の言い換えを生成する行為



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 (entailment)

■ 必ずしも同じ意味とは限らない

- $X \rightarrow Y$: Y holds if X holds

Mouton & Co. is the publisher that published
Noam Chomsky's *Syntactic Structures* in 1957.

The author of *Syntactic Structures* is Noam Chomsky.

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

march → walk Troponymy

forget → know Backward presupposition

has started → started Temporal

- 言い換え: 両方向の含意 [Dagan+, 06]

推論 (inference)

■ 真理値の同一性は保証されない

- 発語内行為と類似

Mouton & Co. is the publisher that published
Noam Chomsky's *Syntactic Structures* in 1957.

Mouton & Co. benefited from Chomsky's *Syntactic Structures*.

- 他の例

My son eats eggplants.

My son likes eggplants.

Everything is imported to Japan.

Everything is eaten in Japan.

- ただしこの計算が役立つ場合もある [Pantel+, 07]

まとめ

■ 同義性のレベル [Sato, 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

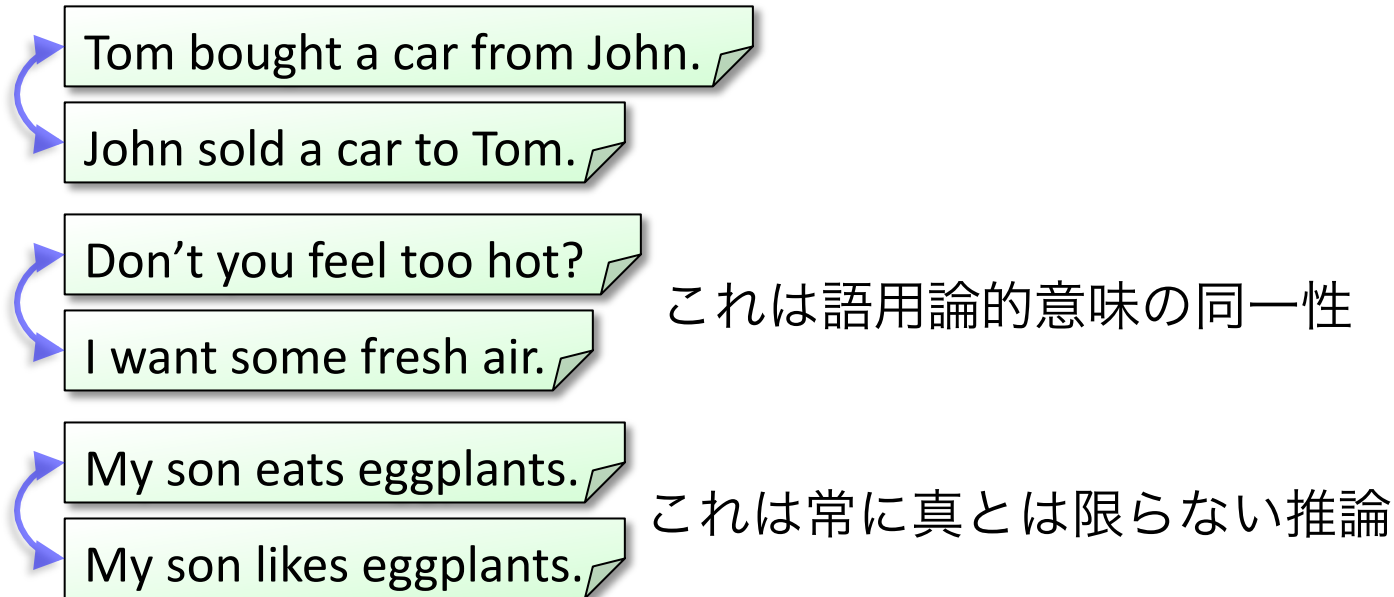
- 含意: 言い換え(両方向の含意) ⊂ あらゆる含意
- 推論: 含意 ⊃ 正しい推論

もくじ

1. 意味の同一性
- ▶ 2. 言い換えの分類体系
3. 計算～知識獲得
4. 言い換え技術の応用例
5. 今後の研究課題

言い換えの分類体系

■ これは言い換え？



■ 分類の意義

- 人間が行っている計算の種類の俯瞰
- 自動化のための基盤
 - 必要なリソースの検討
 - 目的に応じて選択的に使えるようなモジュール化

言い換えの分類体系

■ 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]
- 句レベルのオラクルを8種類に分類 [Max+, 12]
- Lexical, Phrasal, Syntactic [Ganitkevitch+, 13]
- 語彙的な観点での分類 [Bhagat+, to appear]

■ 議論不足の点

- 原子性

Emma burst into tears and he tried to comfort her.

Emma cried, and he tried to console her.

文レベル?

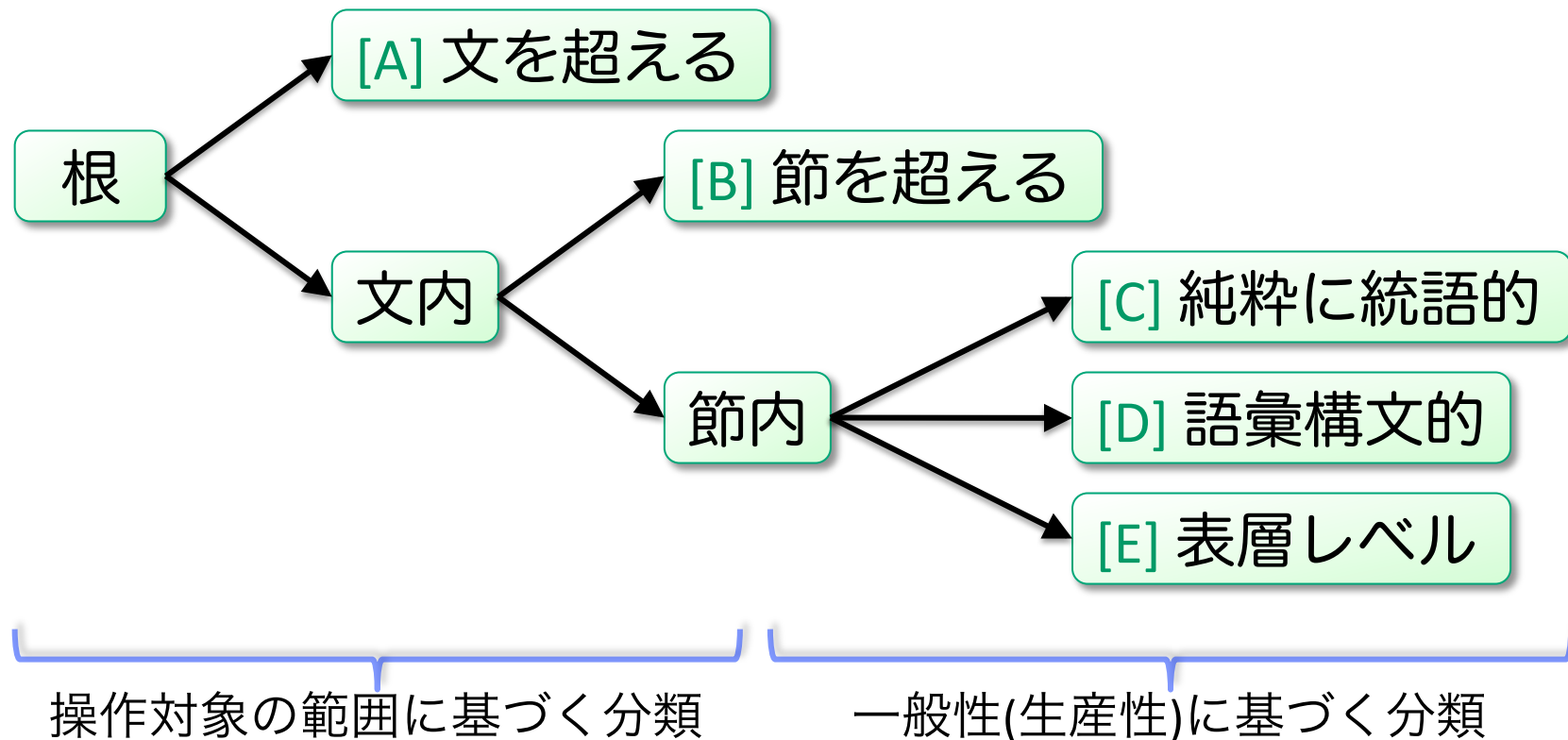
- 網羅性

- おそらくは手法や結果を中心に議論されてきたため

我々の分類

■ 5種類の大分類と60種類の小分類 [Fujita, 10]

- ポスト [乾+, 04]
- ポスト <http://paraphrasing.org/> [藤田, 07]



[A] 文を超える言い換え

■ 複数の節の分割・統合 (関係節)

Småland, **which is** located to the south-west of Stockholm, **is** called “The Kingdom of Glass”. The reason is that there are sixteen glass manufacturers in this area.

Småland **is** located to the south-west of Stockholm. **It is** called “The Kingdom of Glass”. The reason is that there are sixteen glass manufacturers in this area.

[B] 節を超える言い換え

■ 分裂文 ↔ 無標文

It was his best suit that John wore to the dance last night.
John wore his best suit to the dance last night.

■ 主辞の交替 (関係節 ↔ 節の補語)

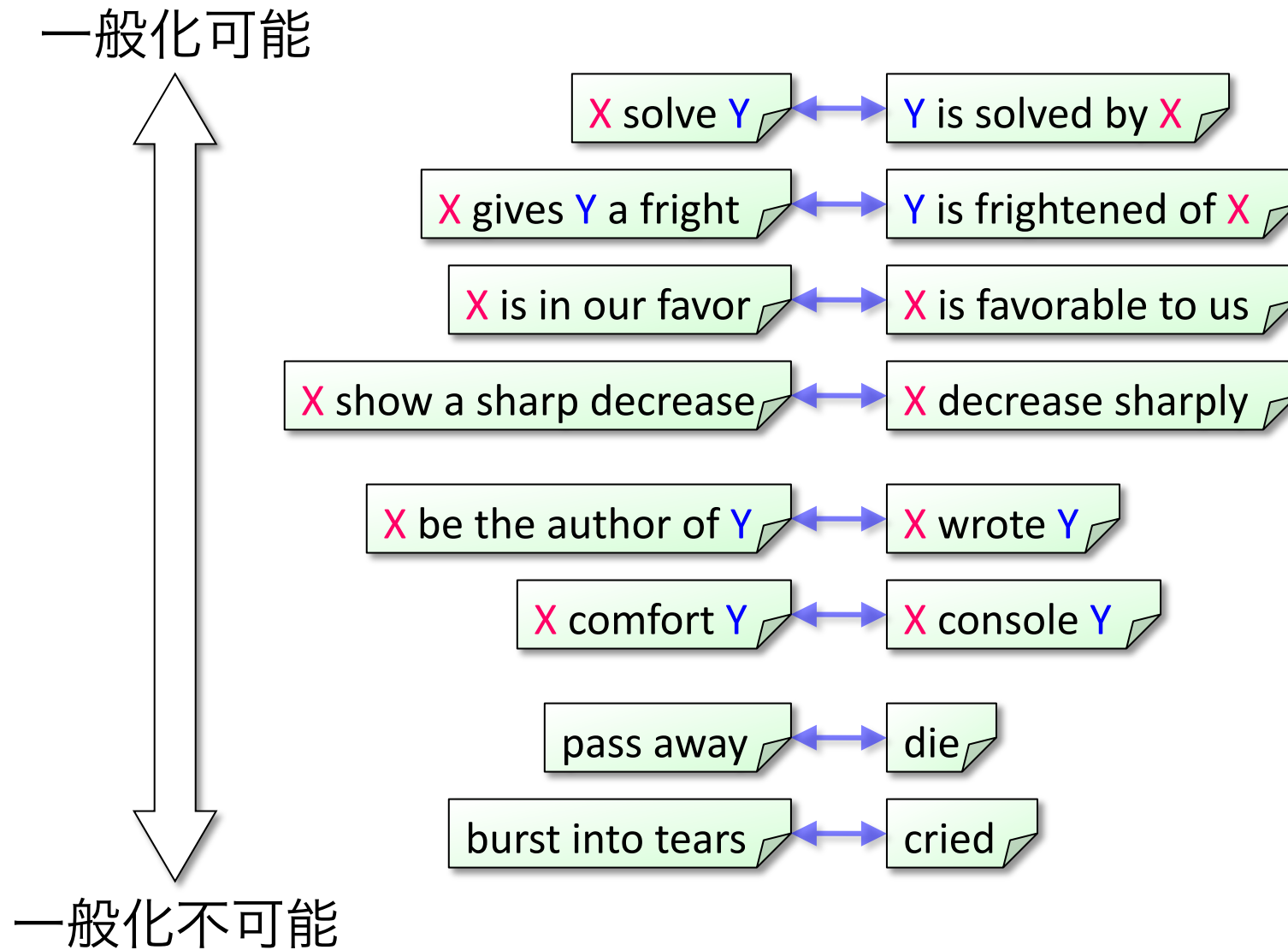
The conference venue is the building whose roof is red.
The conference venue is the building with a red roof.

■ 否定表現の移動

Your application is canceled if you do not reply.
Your application is not canceled if you reply.

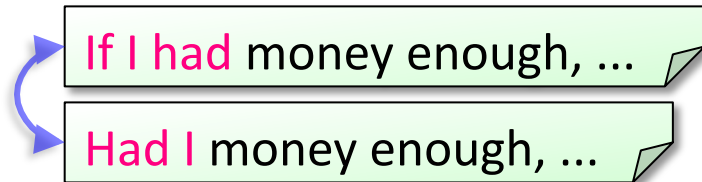
■ 埋め込み節 ↔ 並列節, 並べ替え, etc.

[C, D, E] 単一の節内の言い換え

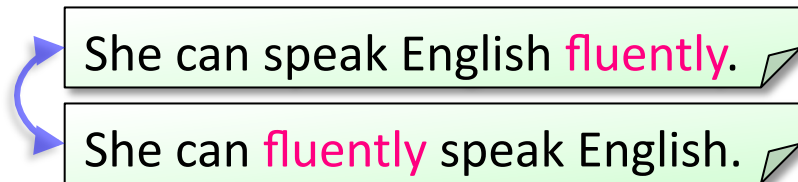


[C] 純粹に統語的な言い換え

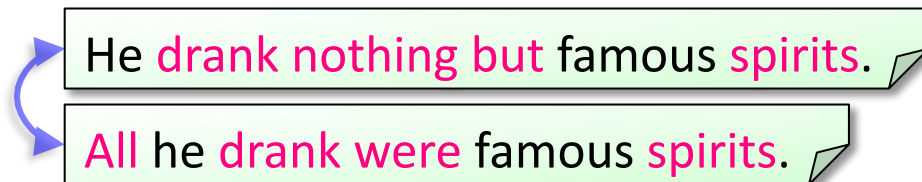
■ 倒置



■ 副詞の移動



■ 否定表現の言い換え



■ あまり多様性はないので書き尽くせる可能性あり

[E] 表層レベルの言い換え

- 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

- 類義語

There's a **risk** of receiving a **severe wound**.

There's a **possibility** of receiving **serious injury**.

名詞, 形容詞

- 慣用表現

Emma **burst into tears** and he tried to **comfort** her.

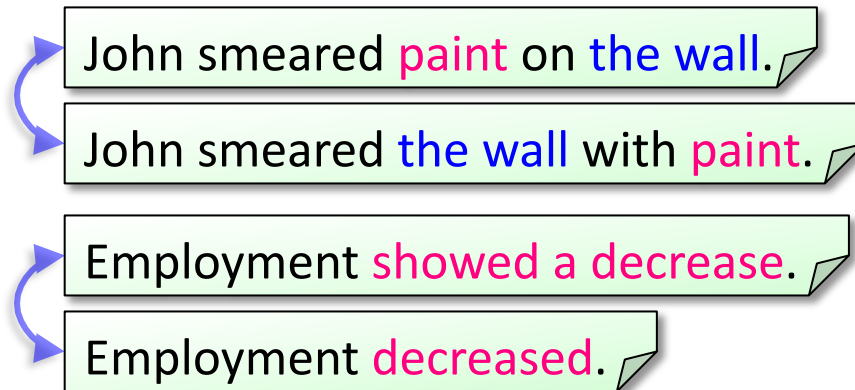
Emma **cried**, and he tried to **console** her.

動詞, 動詞句

[D] 形態構文的な言い換え

■ 一見すると統語的な言い換え[C]のようだが

- ある程度の語彙依存性



- 語彙意味的な情報が不可欠

■ 再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

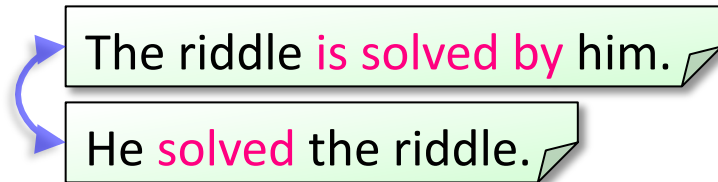
■ 語間の関係に関する情報

- e.g., 語彙的派生関係, 反義関係

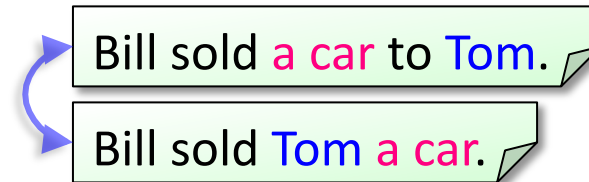
[D]の一種: 動詞交替

[Levin, 93]

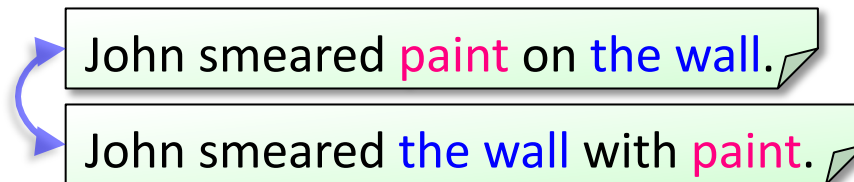
■ 態の交替



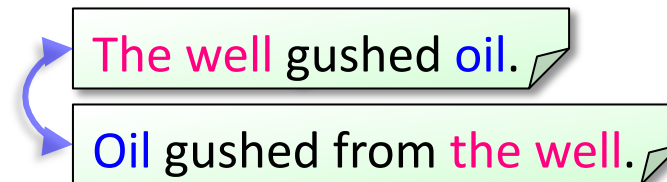
■ 与格交替



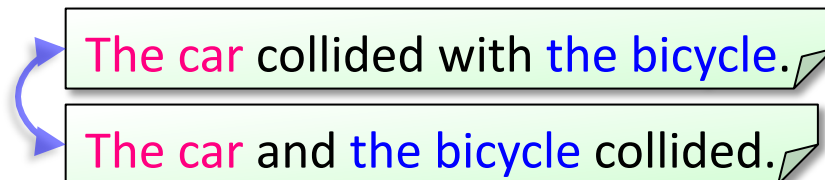
■ 場所格交替



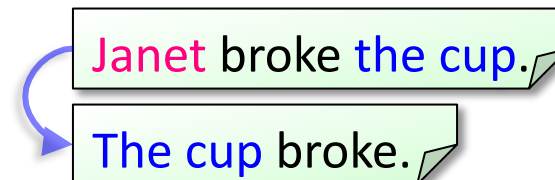
■ 源泉の交替



■ Reciprocal alt.



■ 自他交替 (含意)



[D]の一種: 品詞交替

- 機能動詞構文 (名詞 ↔ 動詞), 形容詞 ↔ 副詞

Employment **showed** a **sharp decrease**.
Employment **decreased sharply**.

- 形容詞 ↔ 動詞

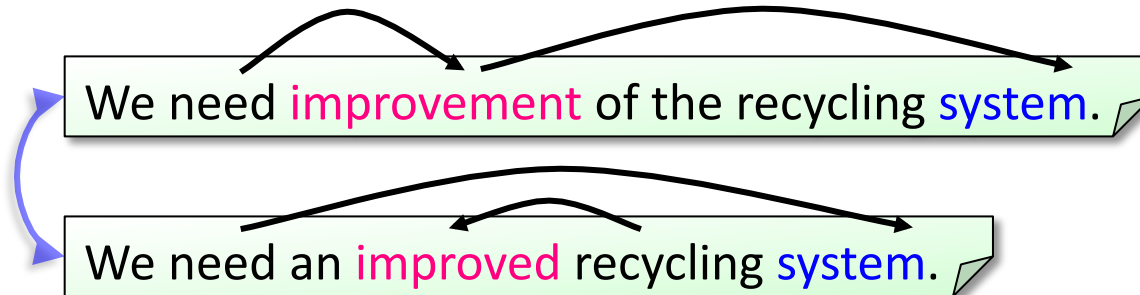
The offer was **fascinating** to everyone there.
The offer **fascinated** everyone there.

- 形容詞 ↔ 名詞

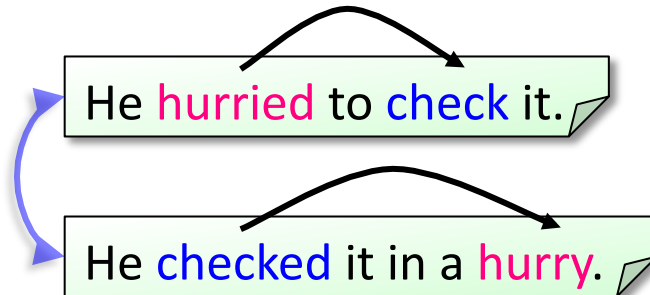
You will not feel **drowsy**.
You can relieve **drowsiness**.

[D]の一種: 構造変換

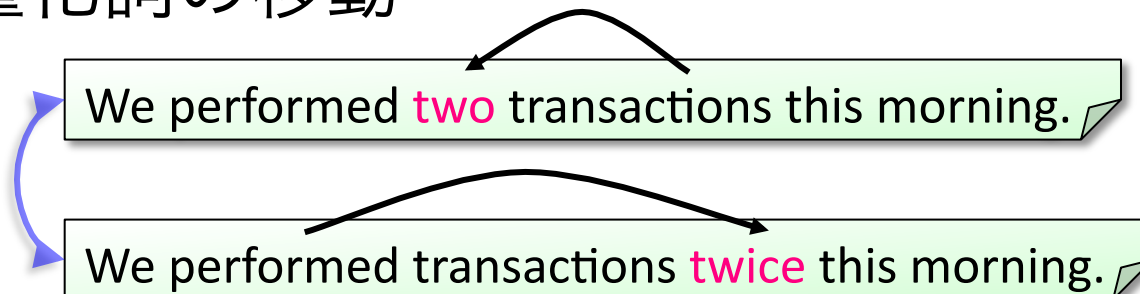
- 主辞の交替 (名詞句), 名詞 ↔ 動詞



- 主辞の交替 (動詞句), 動詞 ↔ 副詞, 名詞 ↔ 動詞



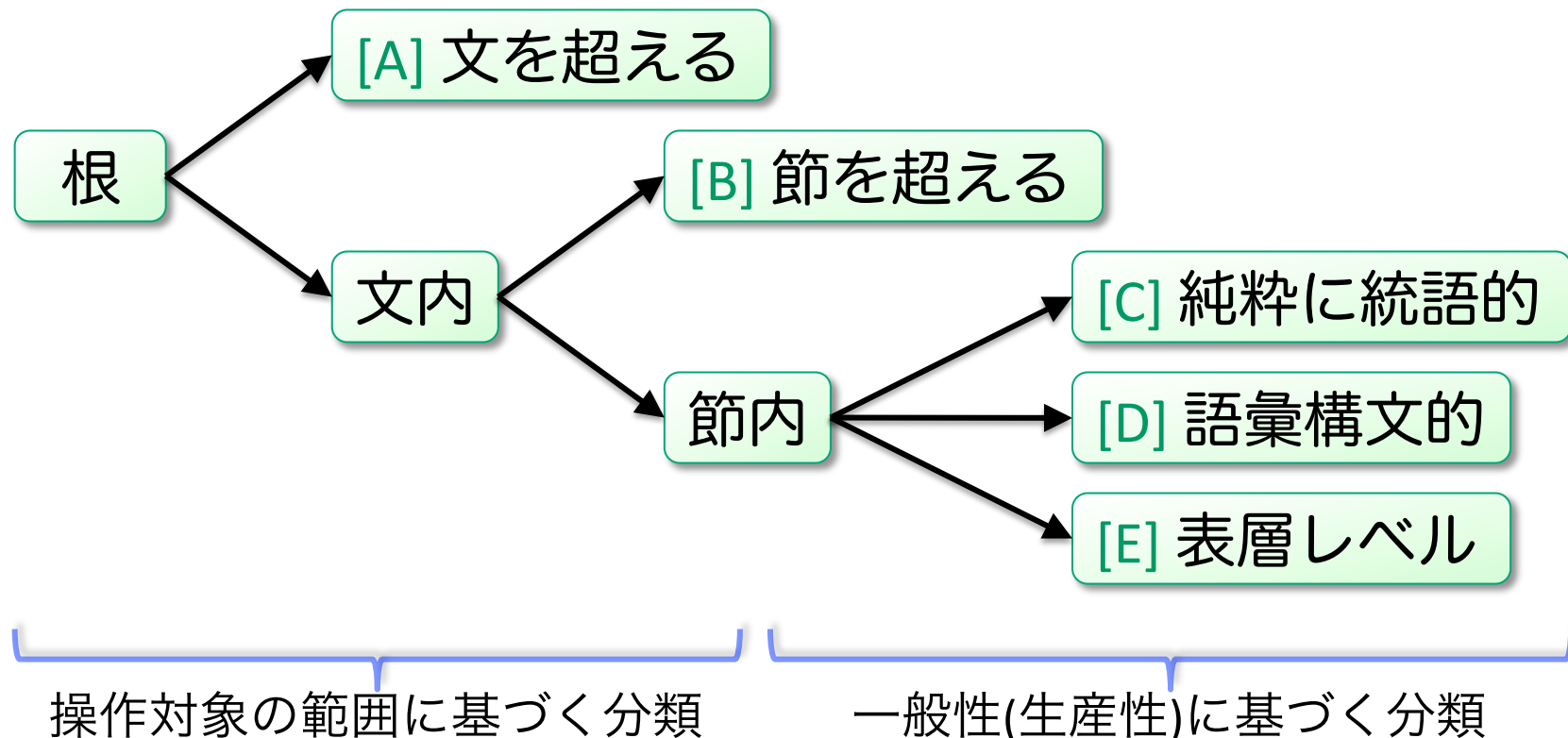
- 量化詞の移動



まとめ

■ 5種類の大分類と60種類の小分類 [Fujita, 10]

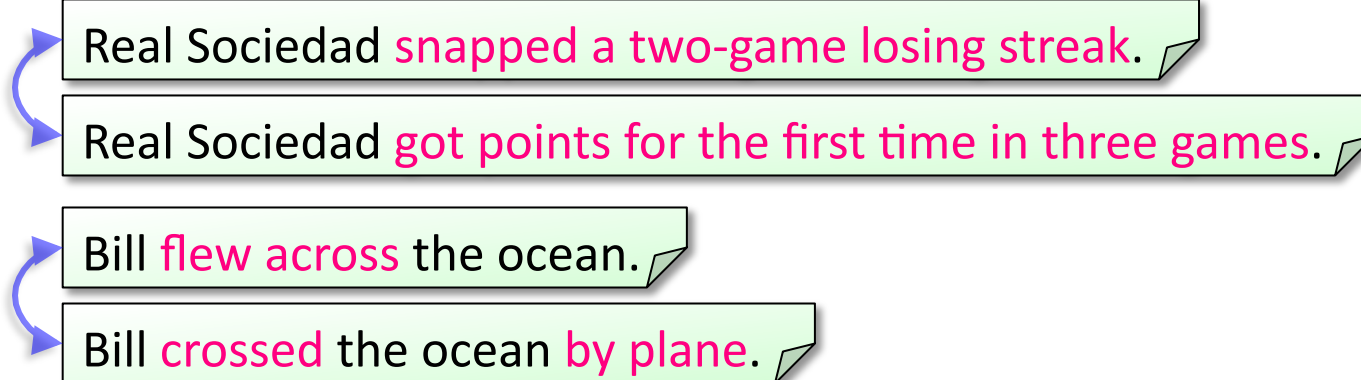
- ポスト [乾+, 04] [藤田, 07]
- 原子性, (網羅性), 操作対象の範囲, 一般性(生産性)
- TODO: 必要な言語資源, 影響の及ぶ範囲, 差分の分類



議論 (1/2)

■ 分類体系

- 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
- 適切に分類できない例の扱い



議論 (2/2)

■ どうやって実現するか

- [A-C]
 - 種類は少ない
 - 過剰生成が許されるならば書き尽くしてはどうか
- [D] 統語と語彙の相互作用を要解明 → 語彙意味論
 - Meaning-Text Theory [Meřčuk+, 87]
 - 語彙概念構造 [Jackendoff, 91]
 - 生成語彙論 [Pustejovsky, 95]
 - FrameNet [Baker+, 98]
 - VerbNet [Kipper+, 00]
- [E] 自動獲得だのみ
 - 既存の語彙資源も最大限に活用

もくじ

1. 意味の同一性
2. 言い換えの分類体系
- ▶ 3. 計算～知識獲得
4. 言い換え技術の応用例
5. 今後の研究課題

言い換え技術の2つの側面

■ 言い換え認識 (Paraphrase recognition)

- 2つの異なる言語表現 $\rightarrow$ ラベル $\in \{=, \neq\}$
 - 文脈の類似度 (分布仮説) [Harris, 64]
 - 構成要素の類似度: e.g., tree-kernel [Collins+, 01]
 - 言い換え生成での再現可能性 [Takahashi+, 03]
- 言い換え知識獲得は特別なケース
 - 同義である文の集合, 原子的な言い換え知識

■ 言い換え生成 (Paraphrase generation)

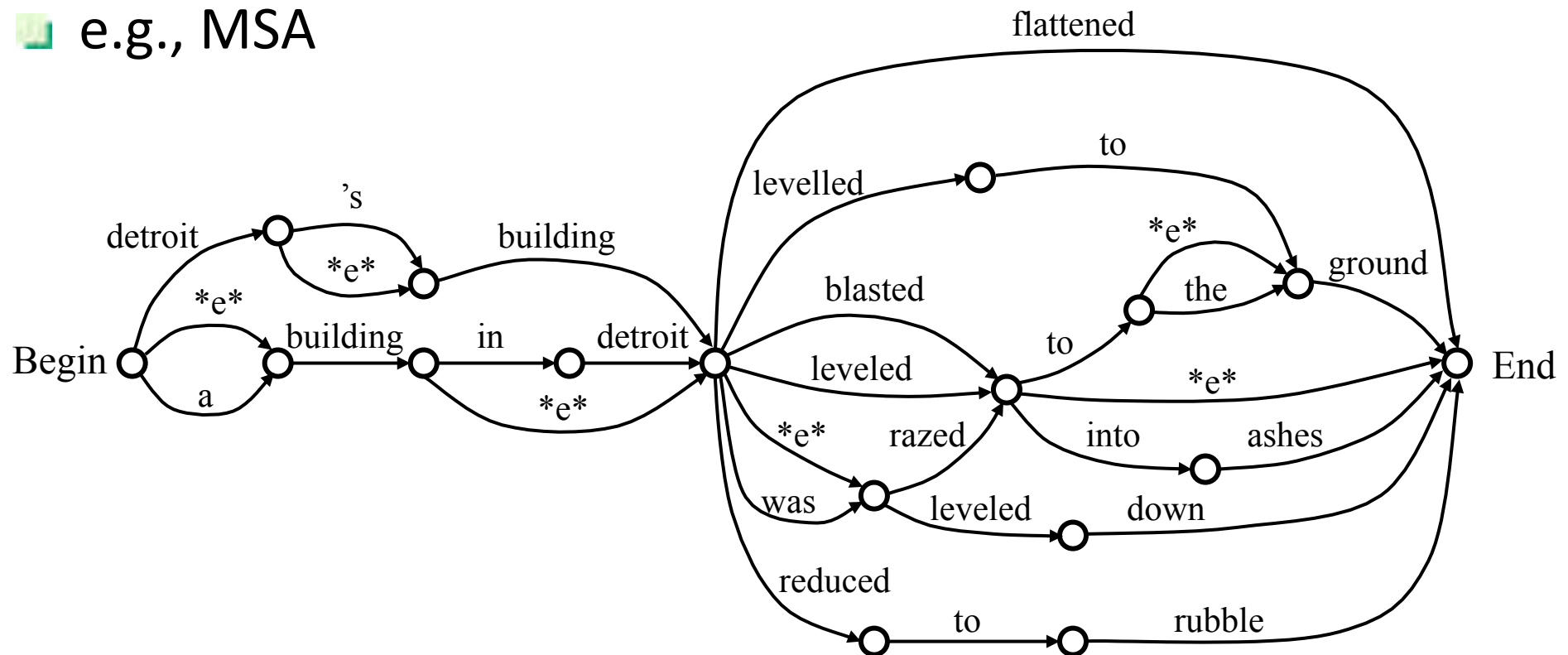
- 目的に敵った(あるいはあらゆる)言い換えを生成
 - 文全体を書き換えるアプローチ
 - 必ずしも文全体を書き換えないアプローチ
- 生成可能 $\rightarrow$ 認識可能 (逆はNG)

文全体を書き換えるアプローチ (1/2)

≡ 単言語機械翻訳

- Multiple-sequence alignment [Barzilay+, 03]
- 重み付きFST [Hori+, 03][Kobus+, 08]
- 統計的機械翻訳 [Quirk+, 04][Wubben+, 10]

e.g., MSA

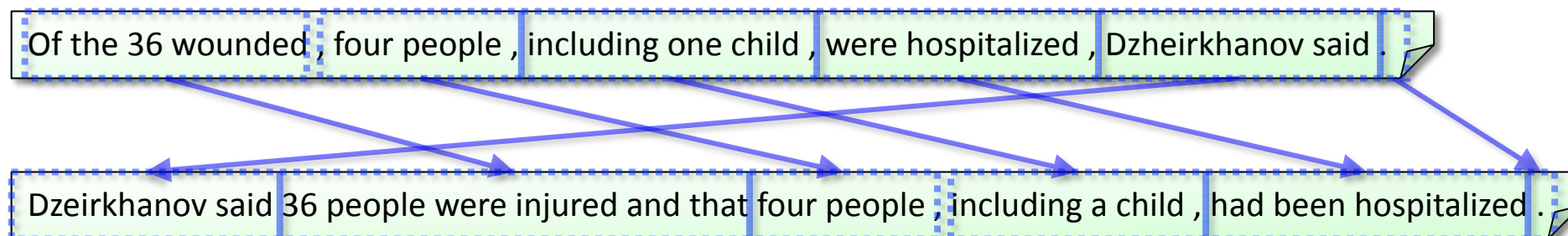


文全体を書き換えるアプローチ (2/2)

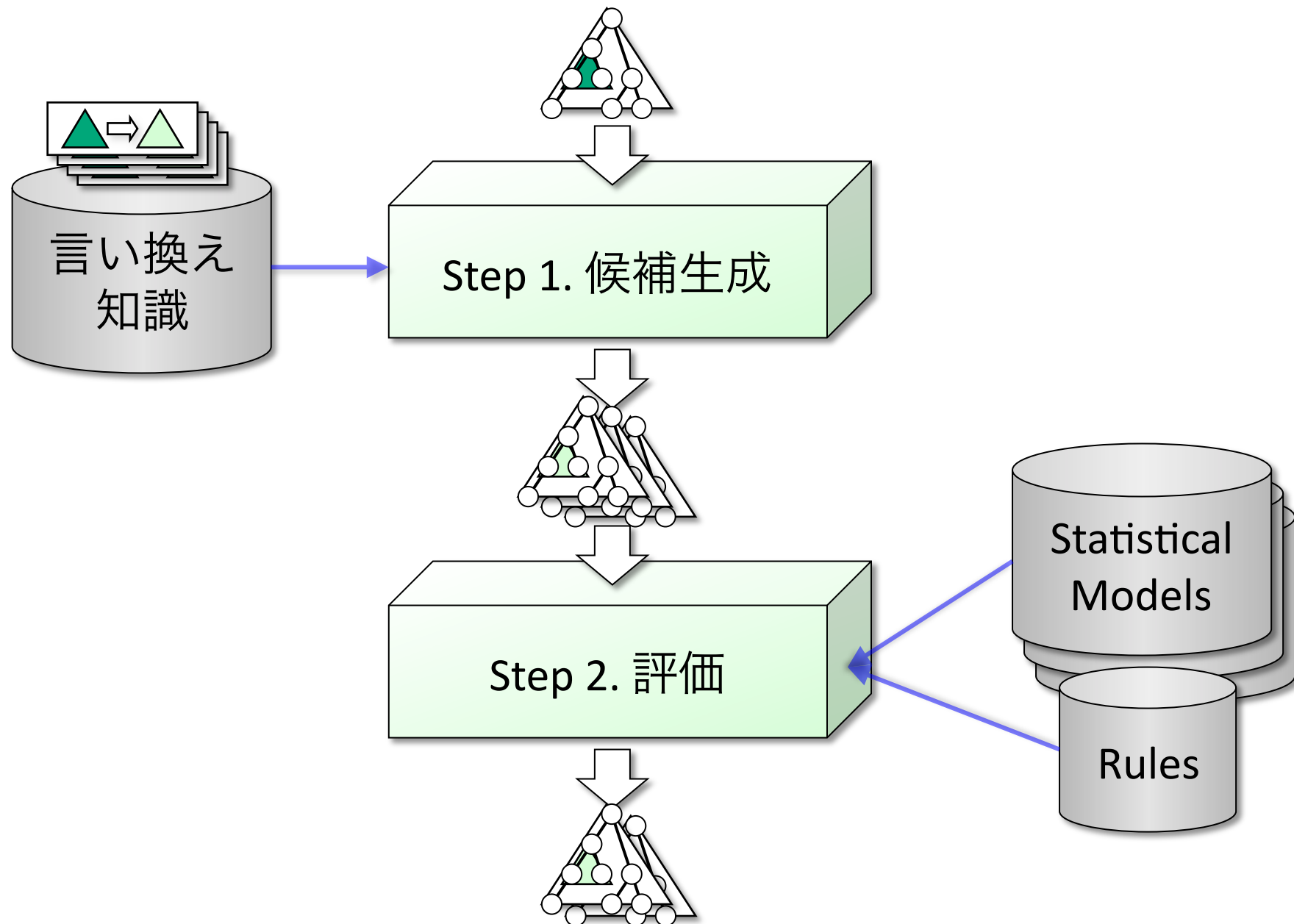
■ ≡ 単言語機械翻訳

- Multiple-sequence alignment [Barzilay+, 03]
- 重み付きFST [Hori+, 03][Kobus+, 08]
- 統計的機械翻訳 [Quirk+, 04][Wubben+, 10]

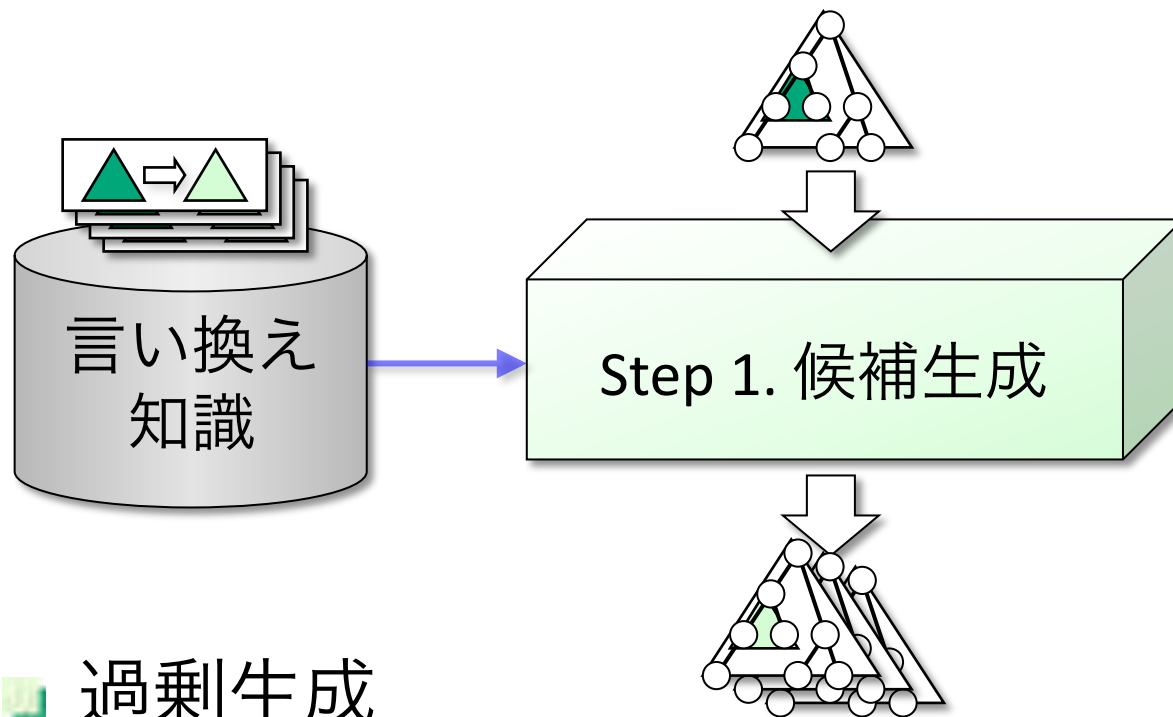
■ e.g., 句に基づく統計的機械翻訳



必ずしも文全体を書き換ええないアプローチ



Step 1. 候補生成

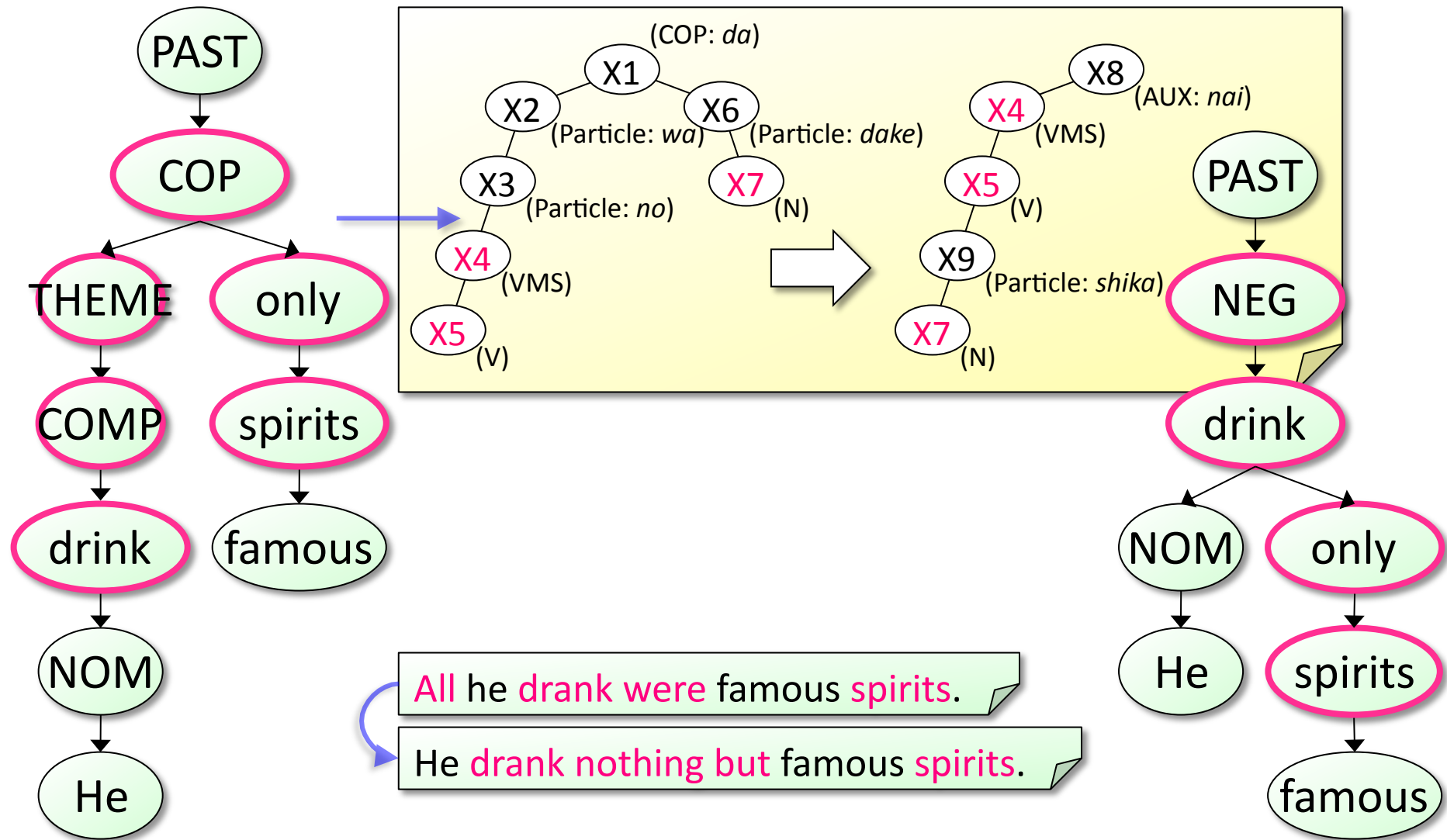


■ 過剰生成

- 入力の一部を置換
- 既存の手法は浅い統語レベル以上を仮定
[Lavoie+, 00][Takahashi+, 01][Zhu+, 10]

e.g., 依存構造レベルの構造変換

[Takahashi+, 01]



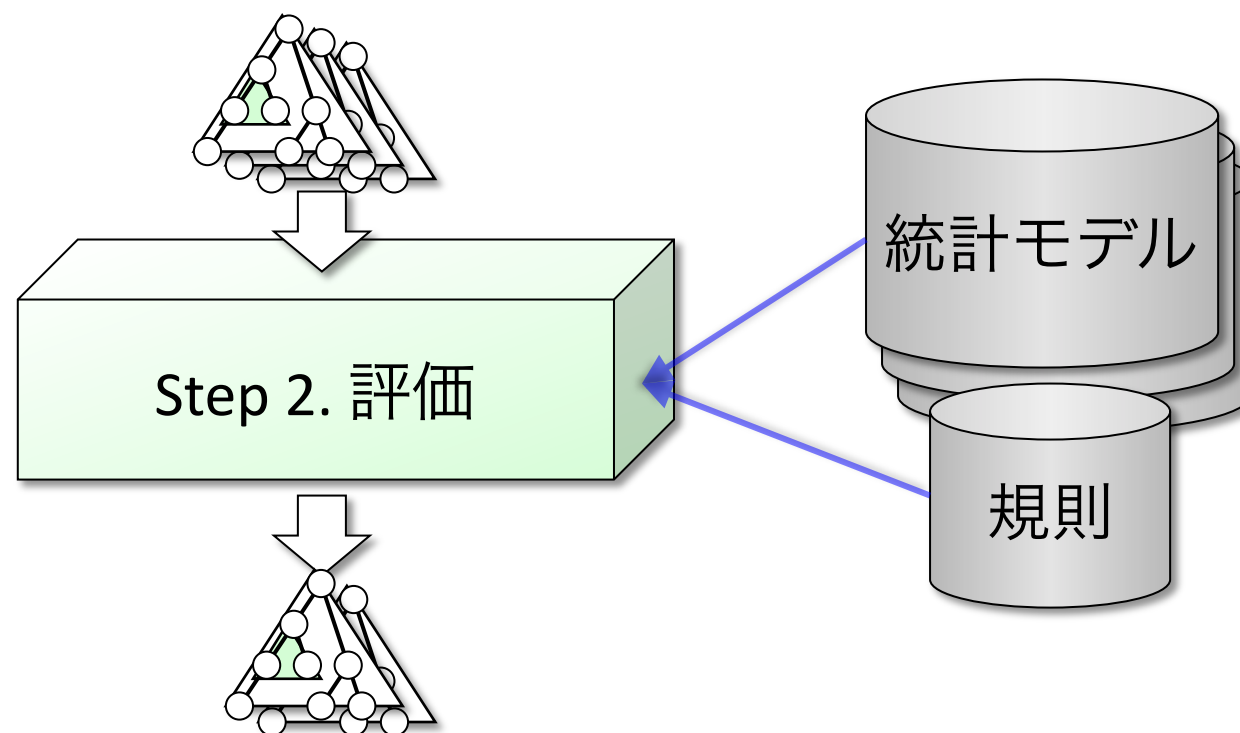
Step 2. 評価

■ 動機

- 入力文の一部のみ書き換えた
- 変換知識は多くの場合十分条件を備えていない

■ フィルタリング, ランキング, 自動訂正

- 規則ベース (語形変化等)
- 統計ベース



統計的な評価

■ 文法性と意味的な適切さ

- 文全体の文法性 [Wan, 05]
- 選択選好 [Fujita+, 04][Pantel+, 07]
 - e.g., <V, relation, N>

■ 文脈における同義性

- 語釈文と文脈の照合 [Okamoto+, 03]
- 選択選好 + トピックモデル [Szpektor+, 08]

比較

■ 文全体を書き換えるアプローチ

- Pro. フレームワークの数学的背景
- Pro. 既存のツールを援用可能
- Con. 特定の表現, 要素を書き換えるための制御

■ 必ずしも文全体を書き換えないアプローチ

- Pro. 動作原理が単純
- Pro. 個々の種類の事例研究にも有用
- Con. 複数の言い換えを組み合わせ制御

■ 共通の課題

- 候補生成の前提となる知識の準備
- 意味論等の深い知識の導入方法

研究課題: 言い換え知識の獲得・蓄積

■ 頑健な処理には大規模な知識ベースが不可欠

- 人手でコツコツ作成
- 自動獲得

■ 各種言い換え知識獲得手法の俯瞰

- 単言語コーパス + 分布仮説
 - 大きい → (相対的には) カバレッジが高い
 - 根拠の信頼性に乏しい → 精度が低い
- n言語パラレルコーパス + 翻訳テーブル
 - 文レベルの同義性が保証されている → 精度が高い
 - データは小さい → カバレッジが低い

手書きの規則, パターン

[A-D]

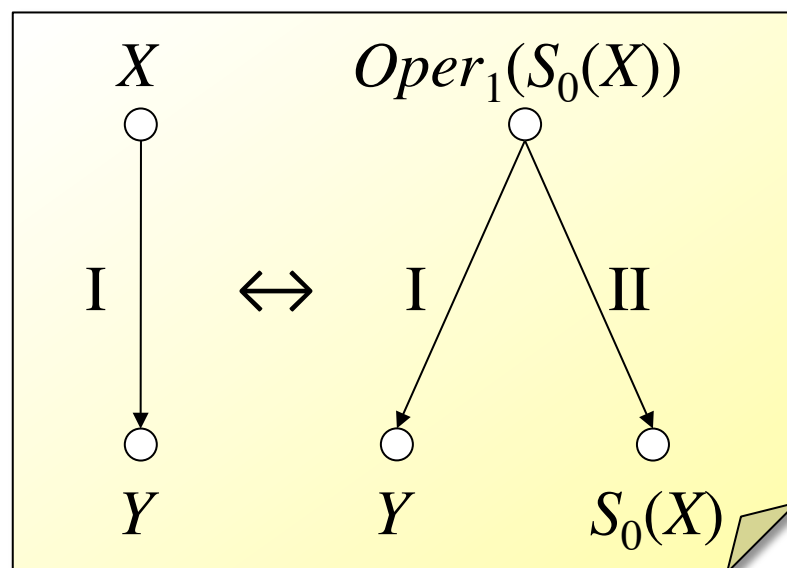
■ フレームワーク

- 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

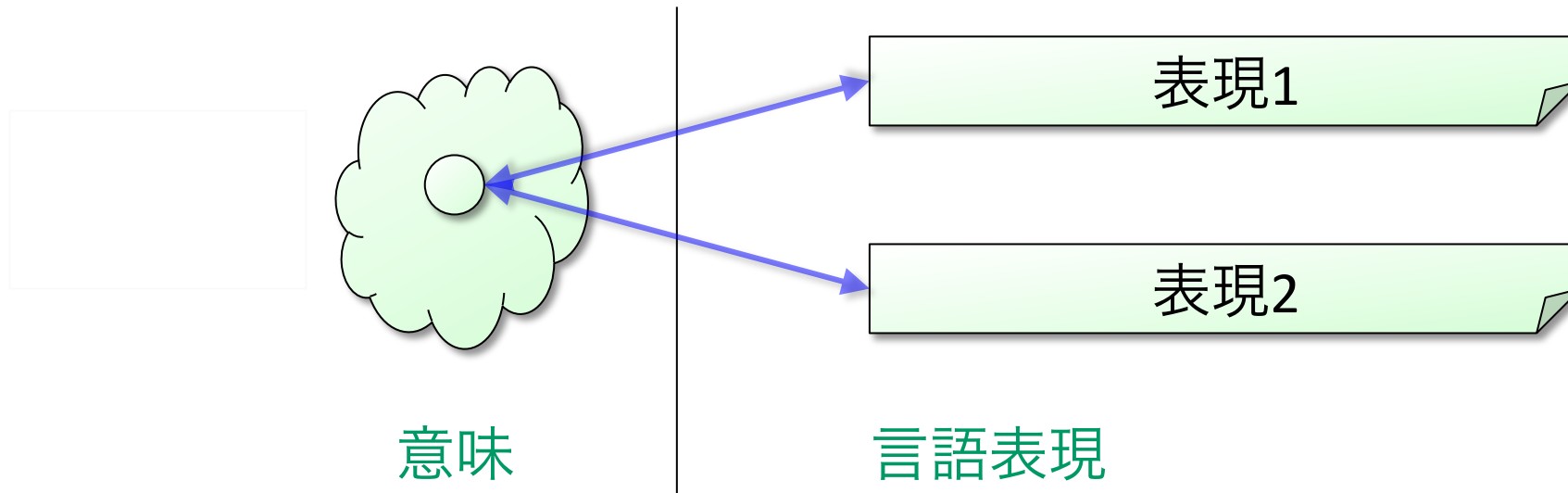
- Meaning-text Theory [Mel'čuk+, 87]



動詞句 ↔
機能動詞構文

言い換え知識の自動獲得

- 言い換え認識の特殊な版
- 同義性を裏付ける(近似する)アンカー



- e.g.,
 - 概念: Synset ID, 辞書の語釈文
 - 文脈: bag-of-(周辺の表現), bag-of-(句レベルの対訳)
 - 情報源: 複数の人間訳, 異なる通信社の関連記事
 - etc.

概念: シソーラス

[E]

■ 同じ概念IDを持つ語句

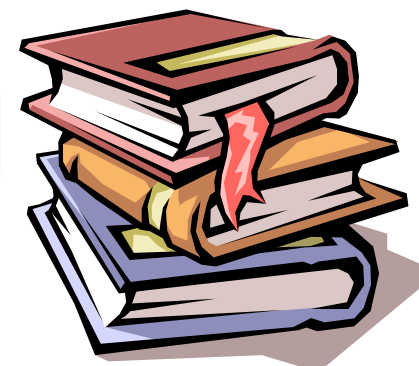
- e.g., WordNet [Miller+, 85]

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



achieve $\leftrightarrow$ accomplish

base $\leftrightarrow$ basis



● 課題

- near-synonym に過ぎない [Clark, 92][Edmonds, 99]
- 多義性や文法性を無視した同義性に過ぎない [Fujita+, 00]
- 継続的な人手が必要
 - e.g., 新語, 固有表現

google_(v) $\leftrightarrow$ search the Web using Google

Future University Hakodate $\leftrightarrow$ FUN

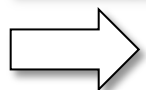
概念: 語釈文

[C-E]

■ 見出し語 ↔ 語釈文

- 各種整形 [藤田+, 00][梶原ら, 13]
- 格フレームアライメント [Kaji+, 02]

compress: make more compact by or as if by pressing



X compress $Y \leftrightarrow X$ make Y more compact

■ 複数の語釈文の対応付け

- 複数の辞書における語釈文 [Murata+, 04]
- Webから得た定義文 [Hashimoto+, 11][Yan+, 13]

Osteoporosis

a disease that decreases the quantity of bone and makes bones fragile

a disease that reduces bone mass and increases the risk of bone fracture

文脈: Bag-of-(周辺の表現)

[B-E]

■ 分布仮説 [Harris, 68]

- “類似する文脈でよく使われる表現は似た意味を持つ”
- 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: e.g., 反義語, 上位語



文脈: Bag-of-(句レベルの対訳)

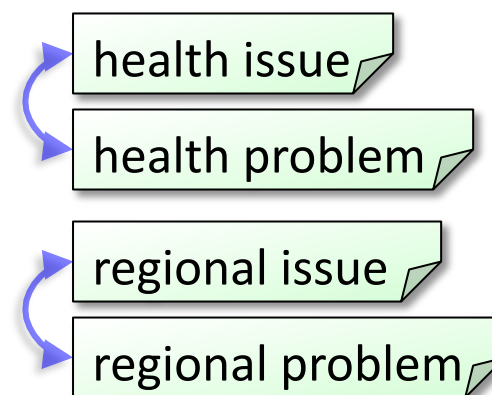
[B-E]

■ 翻訳をピボットとみなす [Bannard+, 05]

- 分布仮説における「文脈」よりは信頼できる
- 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 + 対訳句対の抽出
- 対訳コーパス << 単言語コーパス

自動的に学習した翻訳テーブル

health issue		problème de santé
health problem		problème de santé
regional issue		problème régional
regional problem		problème régional

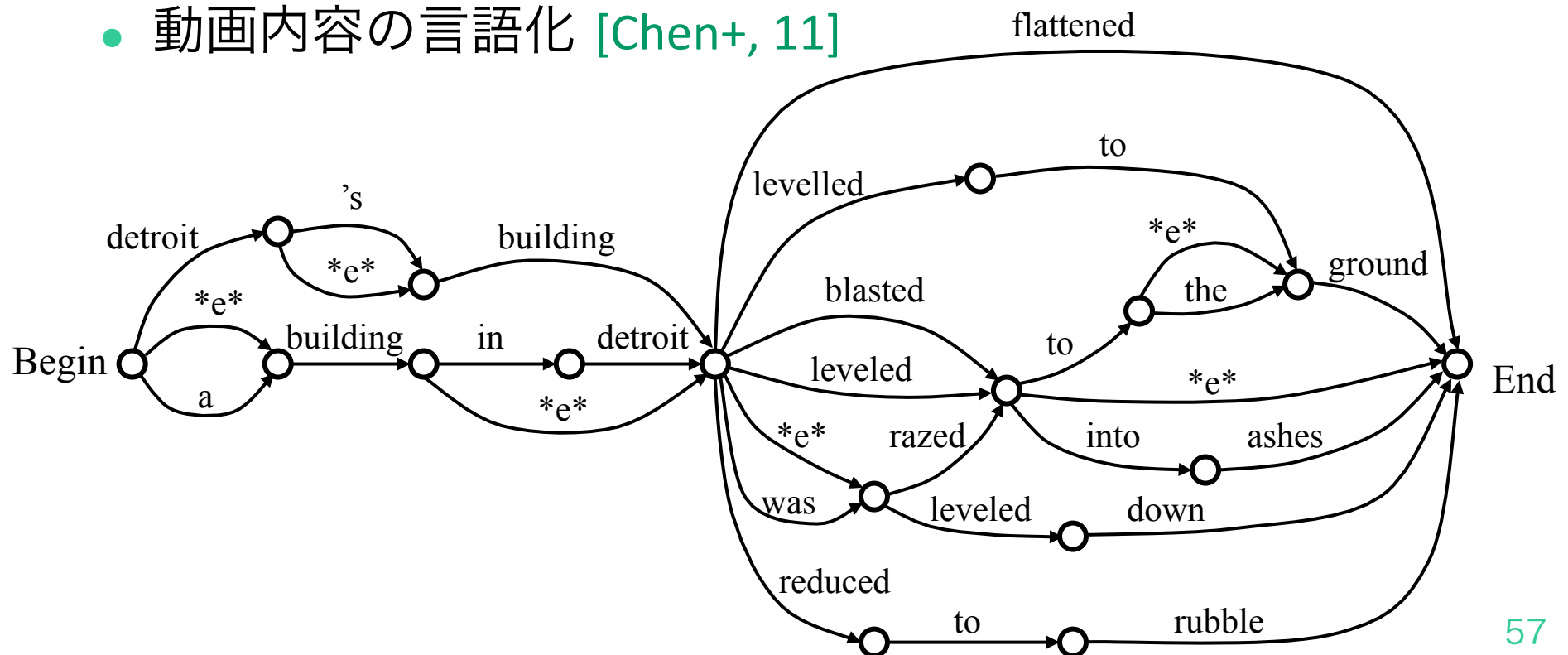


情報源: 異なる言語化

[A-E]

■ 情報源

- 数式の言語化 [Barzilay+, 02]
- 複数の翻訳
 - 小説 [Barzilay+, 01]
 - MT評価用参照訳 [Pang+, 03]
- 動画内容の言語化 [Chen+, 11]



情報源: 複数の記事の対応付け

[A-E]

異なる通信社の記事

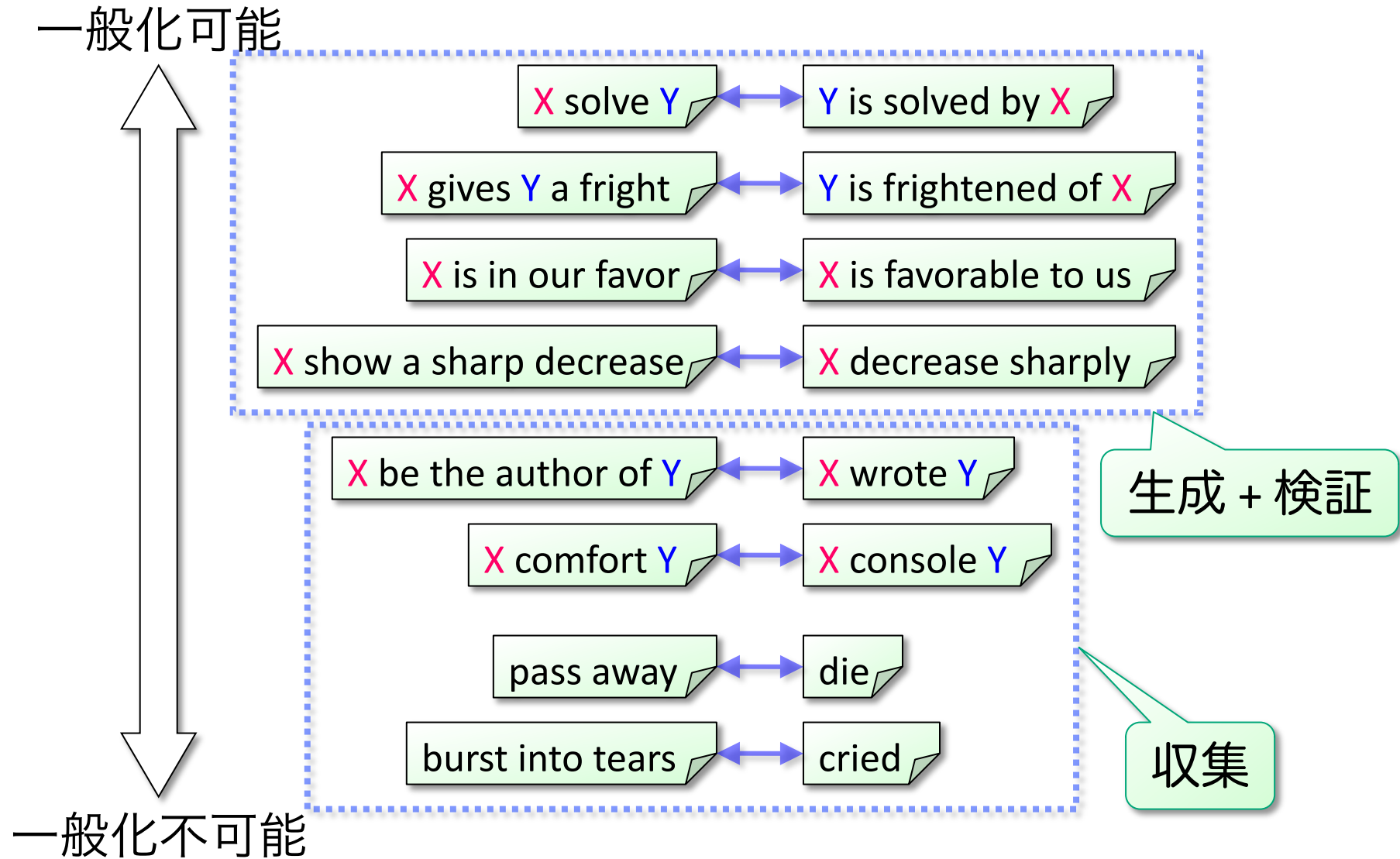
- Web上にたくさんあり, ずっと増え続ける
- パラレルである保証はない → コンパラブル

手法

- 再粒度の固有表現体系 + 分布仮説 [Shinyama+, 02]
- Multiple-sequence alignments [Barzilay+, 03]
- 編集距離に基づくクリーニング [Dolan+, 04]
- Google News の見出し [Wubben+, 09]

単一の節内の言い換えに関する知識

[C-E]

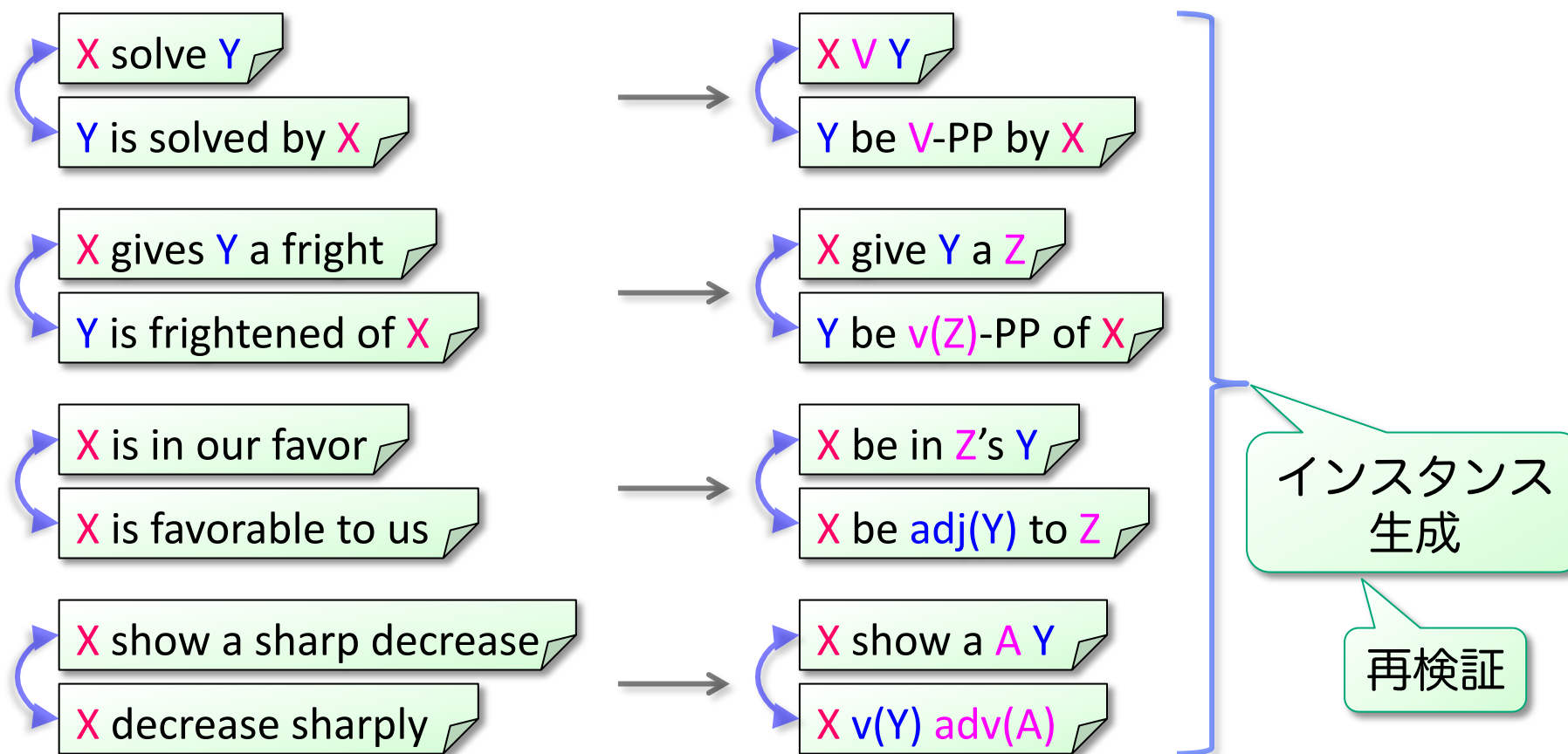


パターンとコーパスの組み合わせ

[D]

■ 事例からの一般化 + コーパスを用いた具体化

- 人間がパターン化 [Jacquemin, 97][Fujita+, 07;08]
- パターンの自動獲得 [Fujita+, 12]



まとめ

課題

- 様々な種類の言い換えのカバー
→ e.g., 分類体系に基づいて1つずつ潰していく

現状

おそらく網羅可能				完全な一般化は困難			
大分類	手書き	自動検出	組み合わせ	大分類	手書き	自動検出	組み合わせ
[A] 文を超える	○	—	—	[A] 文を超える	○	—	—
[B] 節を超える	○	△	—	[B] 節を超える	○	△	—
[C] 純粹に統語的	○	△	—	[C] 純粹に統語的	○	△	—
[D] 語彙構文的	△	△	○	[D] 語彙構文的	△	△	○
[E] 表層レベル	—	○	—	[E] 表層レベル	—	○	—

ここを手書きに頼るのは危ない

Promising?

Promising?

利用可能なリソース

■ 人間の手が入ったリソース

- WordNet [Miller+, 93]
- 日本語WordNet [NICT, 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]
- 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]

■ 自動獲得された大規模データ

- ALAGIN日本語パターン言い換えDB [De Saeger+, 09]
- PPDB: CCG制約つき英語言い換えDB [Ganitkevitch+, 13]
- TERp 付属のデータ (英語) [Snover+, 09]
- METEOR 付属のデータ (英語を含む6言語) [Denkowski+, 11]

もくじ

1. 意味の同一性
2. 言い換えの分類体系
3. 計算～知識獲得
- ▶ 4. 言い換え技術の応用例
5. 今後の研究課題

応用が期待される自然言語処理技術

	言い換え生成	言い換え認識
人間向け	文章執筆支援	剽窃検出
	文章読解支援	翻訳メモリ検索
	TTS前処理	テキスト要約
	MT後編集	QA IR
	ステガノグラフィ	
機械処理向け		SMTの拡張 東ロボ!?
	MT前編集	IE DM

人間向け言い換え認識・生成

e.g., 複数文書要約 [Barzilay, 03]

The image displays a screenshot of a web browser window showing multiple news articles about the Syrian troop pullback from Lebanon. The browser window is titled "USATODAY.com - Syrian troops to begin pullback Monday - Mozilla". The address bar shows the URL "http://www.usatoday.com/news/world/2005-03-06-syria-lebanon_x.htm". The main headline is "Syrian troops to begin pullback Monday". The article text states: "BEIRUT (AP) — Lebanese officials said Syrian troops will begin pulling back toward the Syrian border on Monday, as the leader of the Hezbollah militant group called for a demonstration in Beirut to counter weeks of anti-Syrian protests here." Below the main article, there are sections for "RELATED STORIES" and "BEYOND WORDS". The "RELATED STORIES" section includes links to "Syria, Lebanon OK withdrawal", "Syrian troops to begin pullback", "U.S. raps plan", "Assad announces pullout", "Syria getting peer pressure", "Some Syrians want to leave Lebanon", "Demonstrations continue in Lebanon", and "Bush demands Syria withdraw". The "BEYOND WORDS" section includes a "Multimedia" section with a video link "Video: Saudis meet Syrian president" and "Video: Demonstrations in Lebanon". Below this, there is a "Today's Top News Stories" section with links to "U.S. troops, insurgents clash in Iraq", "Foreign minister says car carrying slain Italian agent in Iraq not speeding", "Syria, Lebanon say troop pullback will be completed by March 31", "Bush selects Bolton as new U.N. ambassador", and "Online tax bill due for smokers".

Another browser window is visible in the background, titled "Columbia Newsblaster: Syria ready for Lebanon pullback - Mozilla". The address bar shows the URL "http://www1.cs.columbia.edu/nlp/newsblaster/summaries/2005-03-07". The main headline is "Syria ready for Lebanon pullback". The article text states: "Lebanese have poured into the streets to protest Syria's military presence in their small Mediterranean country. (article 19) Syrian troops will begin pulling back to the Lebanese border following a Monday meeting of the countries' leaders." Below the main article, there is a "Summary of the United Kingdom, from articles in English" section.

A large white arrow points from the left browser window to the right browser window, indicating a flow or relationship between the two documents.

人間向け言い換え認識

e.g., 情報検索 [Shibata+, 08]

情報爆発プロジェクト 検索エンジン基盤 TSUBAKI

http://tsubaki.ixnlp.nii.ac.jp/index.cgi?start=1&query=かぜ薬を服用するときの留意点

情報爆発プロジェクト 検索エンジン基盤
TSUBAKI

かぜ薬を服用するときの留意点 検索する クリア

かぜ薬を服用するときの留意点 で検索された文書 23 件 検索時間: 2.1 (秒)

- 1 おでびびはうすへ。
id=039078409, score=76.3
母乳 ...授乳中の 服薬
を除き、風邪 薬 程度
おきます。
<http://odevivi.com>
- 2 妊娠とおたふくかぜ
id=026834563, score=76.3
メイン | 市販のかぜ薬は女王かぜ
妊娠とおたふくかぜ 妊娠初期におたふくかぜにかかると、流産の確率が高くなると言われています ...おたふく
かぜによる胎児の先天性異常について、動物実験では発症例はありますが、人についてはなく、おたふくかぜが直接の原因となる先天性異常は
ないと言われています。
<http://ikuji.homn.net/2>
- 3 MDB KW INDEX
id=085895832, score=76.3
・工業用化学品 ... キーワード
す。紹介 ... 【企業・団体事例】
市場動向】物流の構造改革および
<http://www2.mdb-net.com>
- 4 会社情報-佐藤製薬
id=015594144, score=72.4
製薬会社、佐藤製薬の Web
<http://www.sato-seiyaku.com>
- 5 星和書店/季刊精神科臨床サービス/第2巻4号
<http://tsubaki.ixnlp.nii.ac.jp/index.cgi?start=1&query=かぜ薬を服用するときの留意点#>

かぜ ↔ 感冒, 風, 風邪

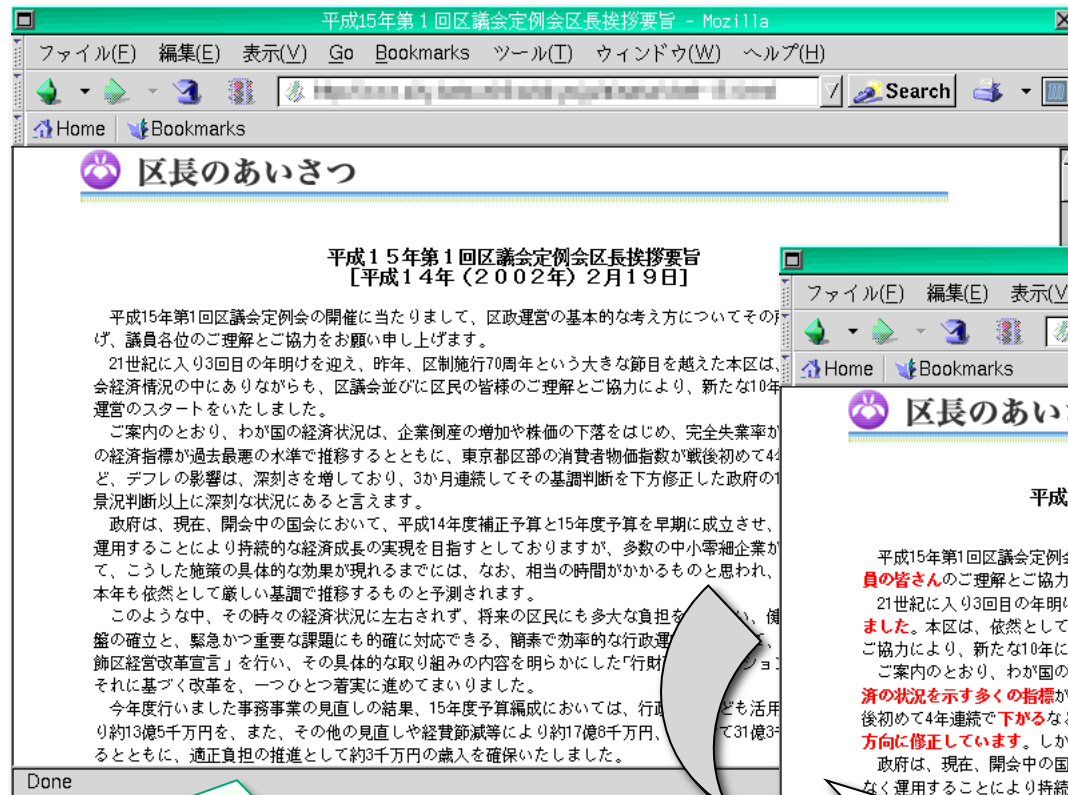
薬を ↔ 薬, 薬品, 薬となる物質, 薬物

服用するときの ↔ 服薬, 服用, 薬を飲む

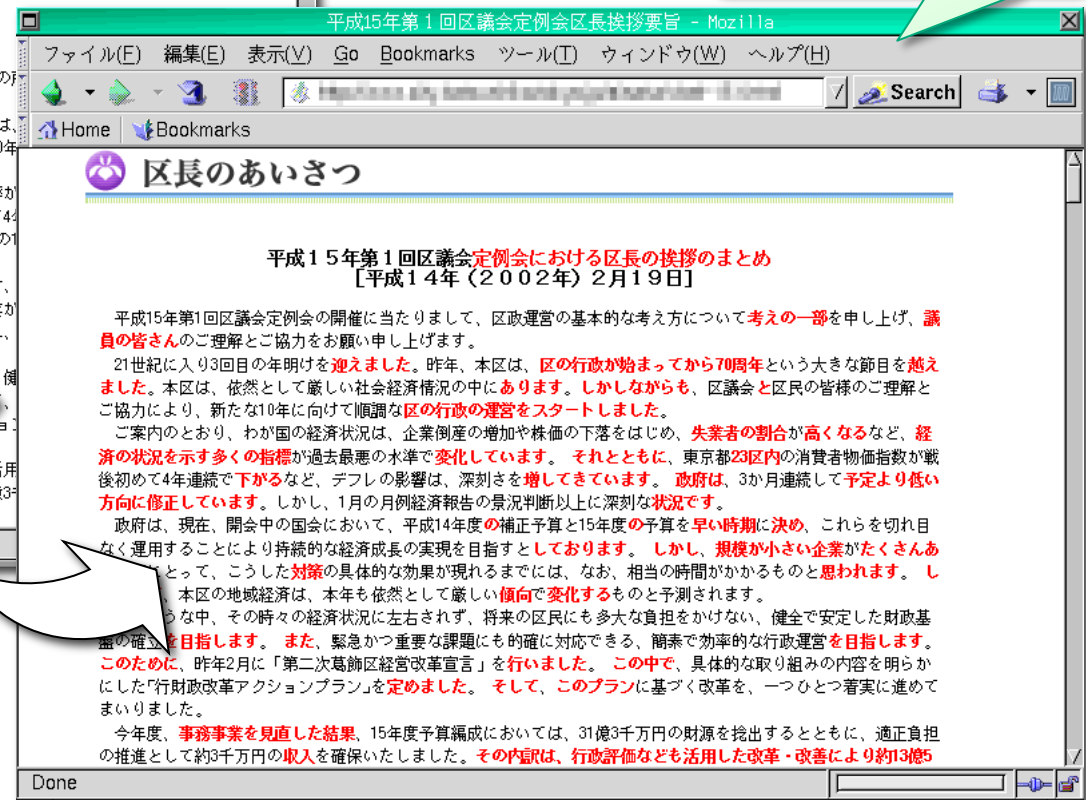
留意 ↔ 気をつける, 用心, 注意, 心に留める

人間向け言い換え生成

e.g., 読解支援のためのテキスト簡単化 [Inui+, 03]

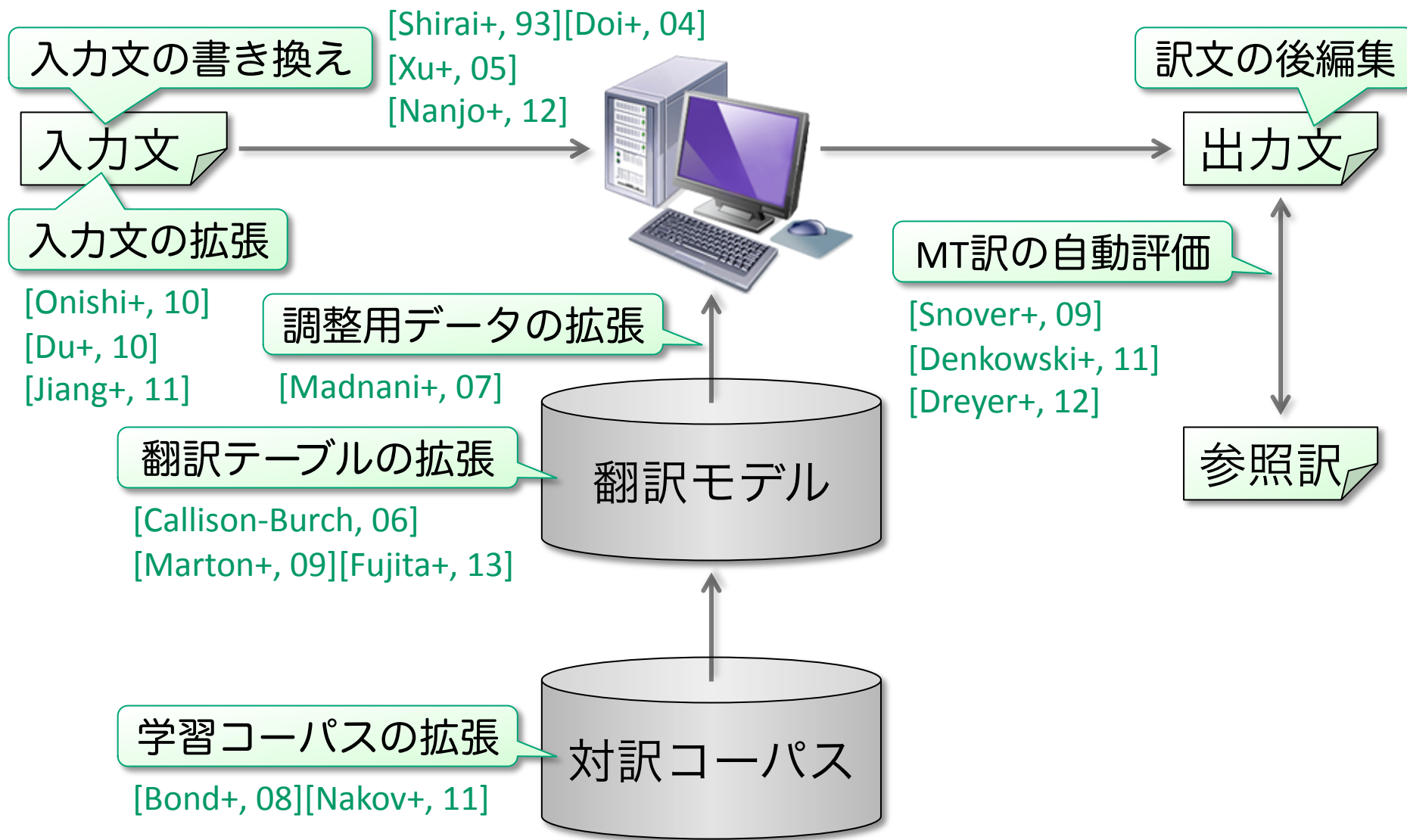


平易, 明瞭



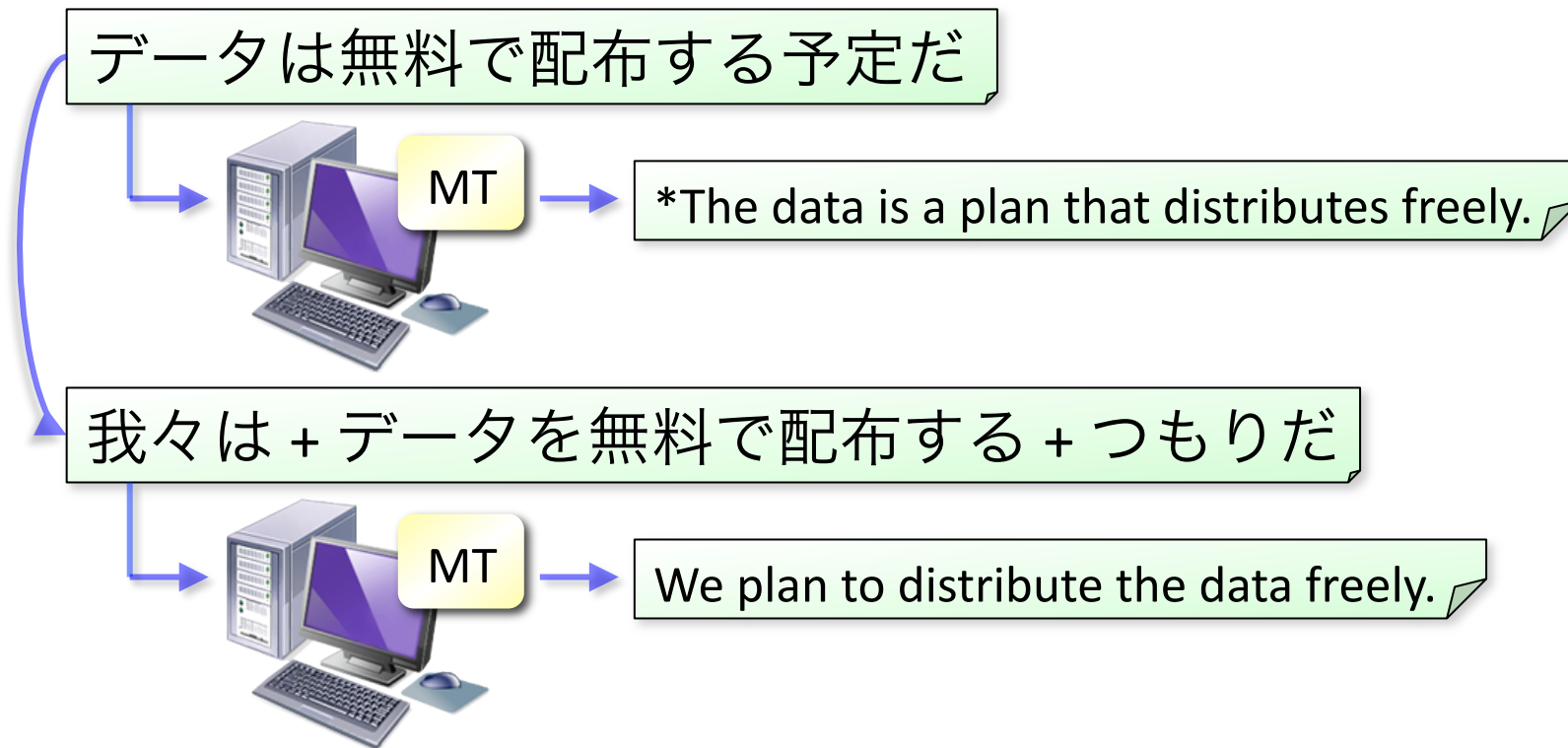
難解, まわりくどい

機械翻訳研究への各種応用



機械処理向け言い換え生成

■ e.g., 翻訳の前編集: 誤りやすい文型の変形 [Shirai+, 93]



- ゼロ代名詞の解析, モダリティ表現の正規化も有用
- 上位語等への言い換え(含意)が必要な場合もある
- 推論してはいけない

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?
あるいはどのような変化は許容できるか
 - 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

1. 意味の同一性
2. 言い換えの分類体系
3. 計算～知識獲得
4. 言い換え技術の応用例
5. 今後の研究課題



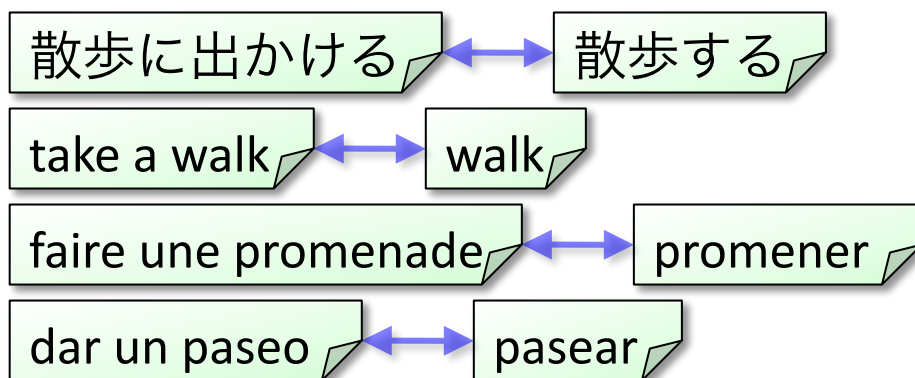
研究課題: 言語学的な観点から

■ 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, コミュニティ作り

■ 言語横断性の分析

- 分類体系の構築とも関連



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

- 知識の形式に関する議論
- 資源の集約
- 仲間を増やして分担
- 効率的な評価手法の確立

■ 知識の適用方法に関する検討

- 言い換え元の表現の多義性解消
- 与えられた文脈における適切さの判定

■ 応用技術に役立てる

- どれくらいの効果があるのかの調査 (oracle)
- 言い換えの種類ごとの選択的な利用

参考文献

- [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+, to appear] Rahul Bhagat and Eduard Hovy. What is a paraphrase? Computational Linguistics, Vol. xx, No. xx, pp.xx-xx, (to appear)
- [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.