

同義の言語表現(言い換え)を 処理する技術の研究動向

藤田 篤



情報通信研究機構

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

自己紹介

■ 藤田 篤 (ふじた あつし)

■ (国研)情報通信研究機構

先進的音声翻訳研究開発推進センター

先進的翻訳翻訳技術研究室 主任研究員 (6年目)

■ 専門: 計算言語学, 自然言語処理

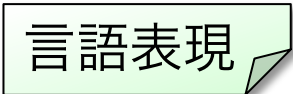
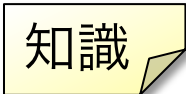
- 自然言語表現の言い換え: FYs 1999-, 21年目
- 翻訳 (機械翻訳, 人間の翻訳): FY 2003, FYs 2011-, 10年目

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:  
- 色塗り文字(, ): 対応箇所, 差分, ハイライト

■ 参考文献の見出しはWeb上の文献リストに対応

- <http://paraphrasing.org/>
- 2019年6月時点で約800件

今回の講演のゴール

■ 会話ができる仲間を増やす

- 情報共有
- そのための「ことば」の共有

■ 研究の広がりや課題について知ってもらおう

- (比較的)言語学的な研究
 - (L1) 定義と境界, 分類体系
 - (L2) 事例研究
- 工学的な研究
 - (L3) 知識やデータの整備
 - (L4) 言い換え計算機構: 言い換え認識, 言い換え生成
 - (L5) 応用技術における活用

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



意味空間



多様性

多義性

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。

言語空間

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

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

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.

■ いずれも目的指向

- 結果物の用途, 言語使用域などを要考慮

言い換えに関する研究の全体像

■ 5つのレイヤ

(L5) 応用技術における活用

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) 応用技術における活用

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

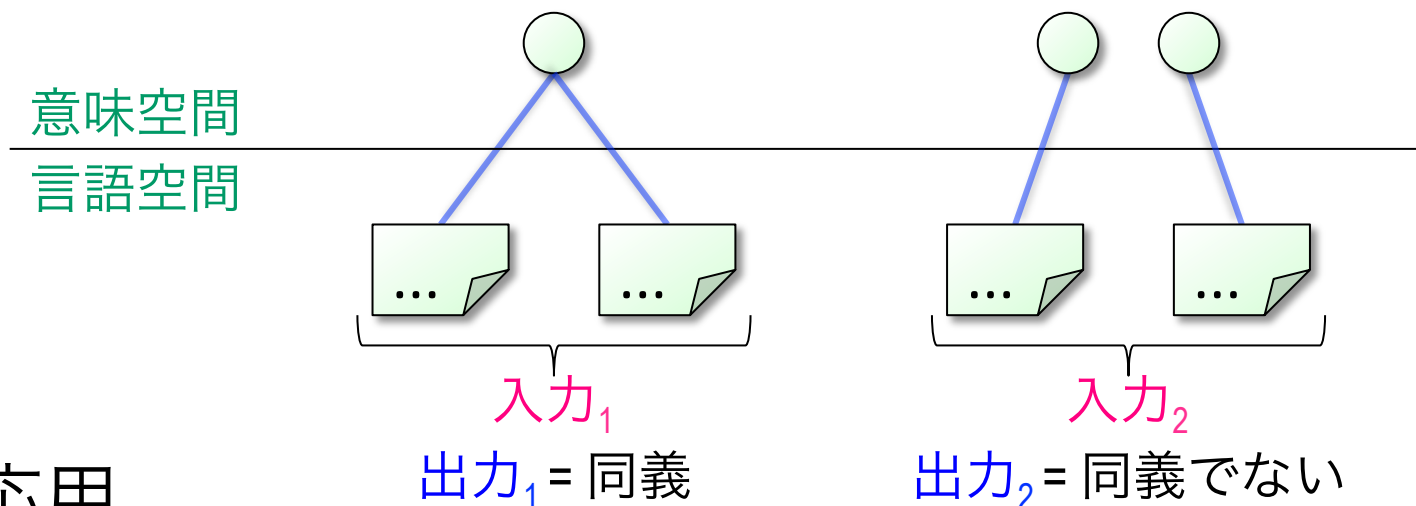
(L1) 分類体系

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



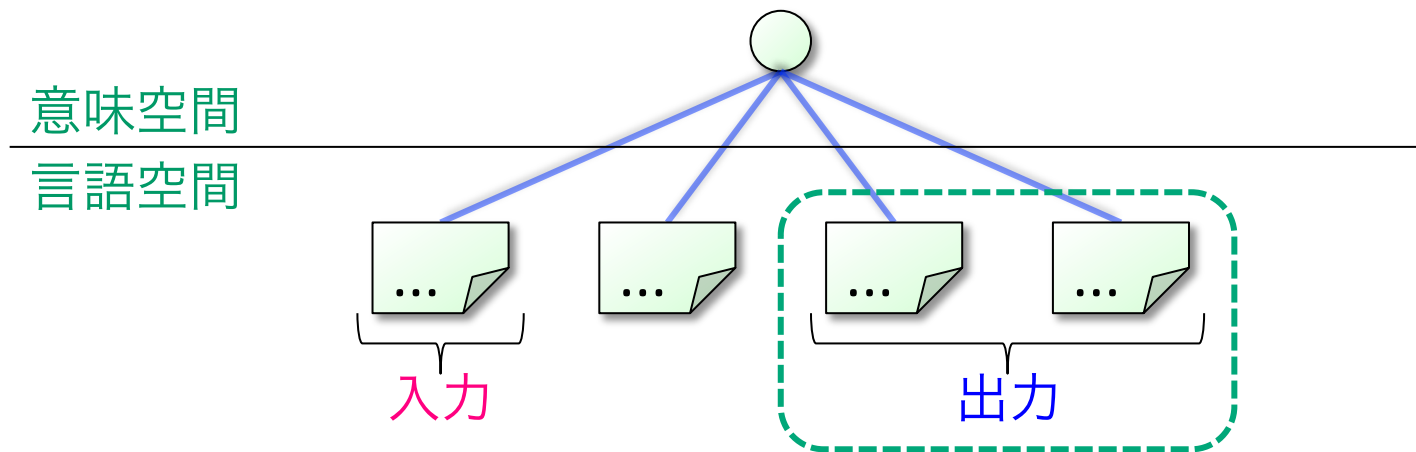
■ 応用

- 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前編集 (e.g., 未知語→既知語)

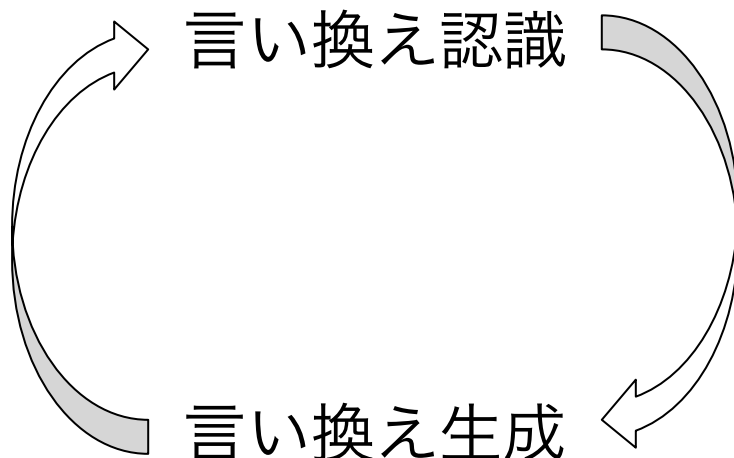
(L4) 表と裏? 卵が先か鶏が先か?

“生成可能ならば正しい言い換え,
できなければ正しくない”

[Takahashi+ 03]

“認識可能ならば探索空間内の候補を
すべて判定することで生成可能”

[Roy+ 19]



■ 2つが近づいてきている例

- Sequence-to-sequence (Encoder-Decoder; EncDec) は認識にも使える生成モデル [Mallinson+ 17][Roy+ 19]
- 認識器を報酬に用いた生成器の強化学習とその生成器から識別器を訓練する逆強化学習 [Li+ 18]

(L3) 知識の整備

■ 言い換えの「説明」に必要な知識

- 言語的な知識

- 語句の素性/意味記述

- e.g., 項構造/意味役割, 動詞ごとの使役や受身の可否

- 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, 上位/下位, 部分/全体, 特質構造

- 世界知識, ドメイン知識

連敗をkで脱出 ↔ (k+1)試合ぶりの勝ち点

森保ジャパンの10番 ↔ 中島翔哉

(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 文書要約	剽窃検出 翻訳メモリ検索 情報検索 質問応答
機械処理向け	ステガノグラフィ MTのモデル拡張 MT前編集 データの拡充	入試問題への解答 データマイニング

(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) 応用技術における活用

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。

暑すぎませんか?

窓を開けてくれませんか?

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。

主題/焦点が異なるが
明示的意味は等しい

語用論的に同じ効果を持ちうる

常に真とは限らない

■ 分類の意義

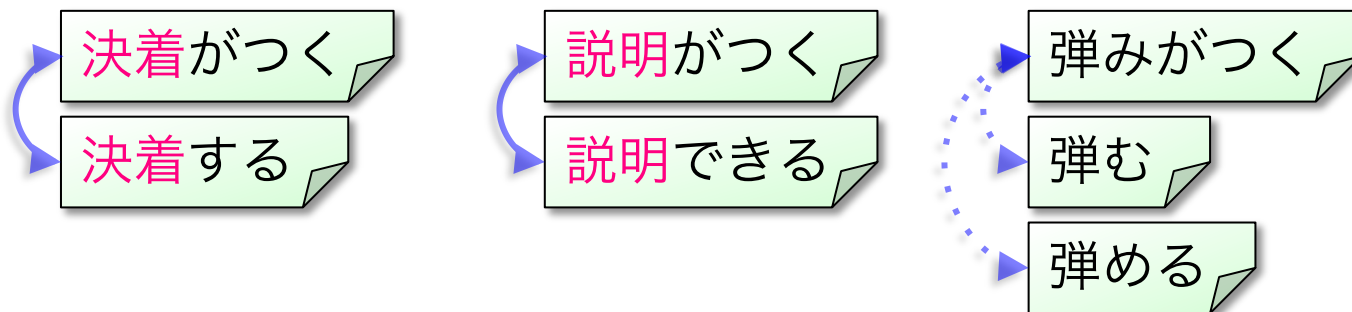
- 人間の言語処理を俯瞰する道標
- 応用の際の取舍選択の判断材料

(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

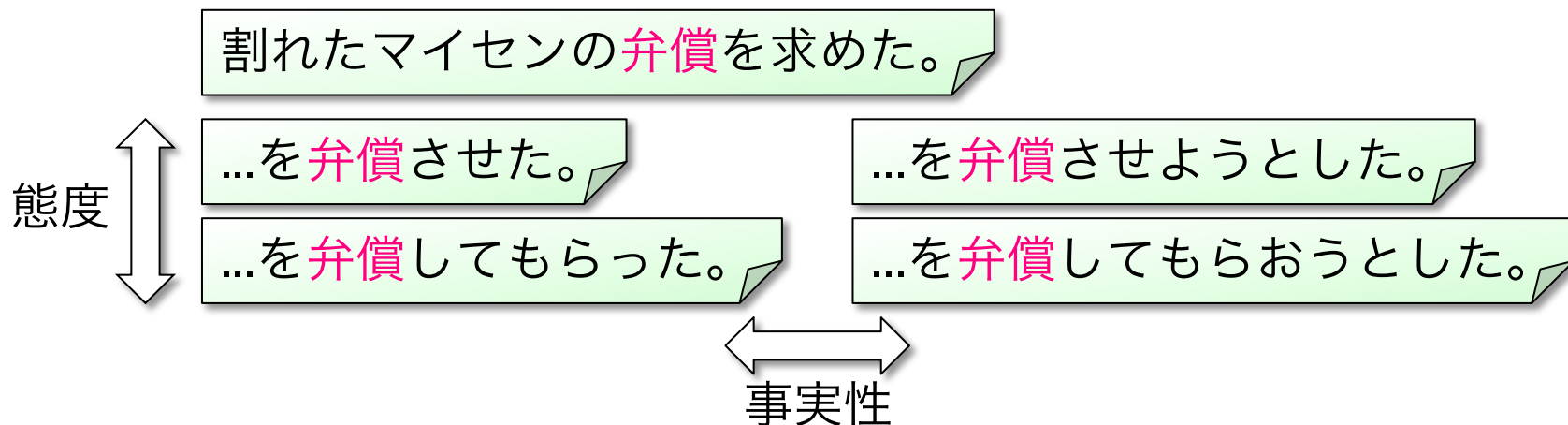
- e.g., 機能動詞構文と動詞句の言い換え

■ 「～がつく」



■ 意味の等価性を厳密に考える題材

- 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基盤

- 情報伝達, 意思疎通
- 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

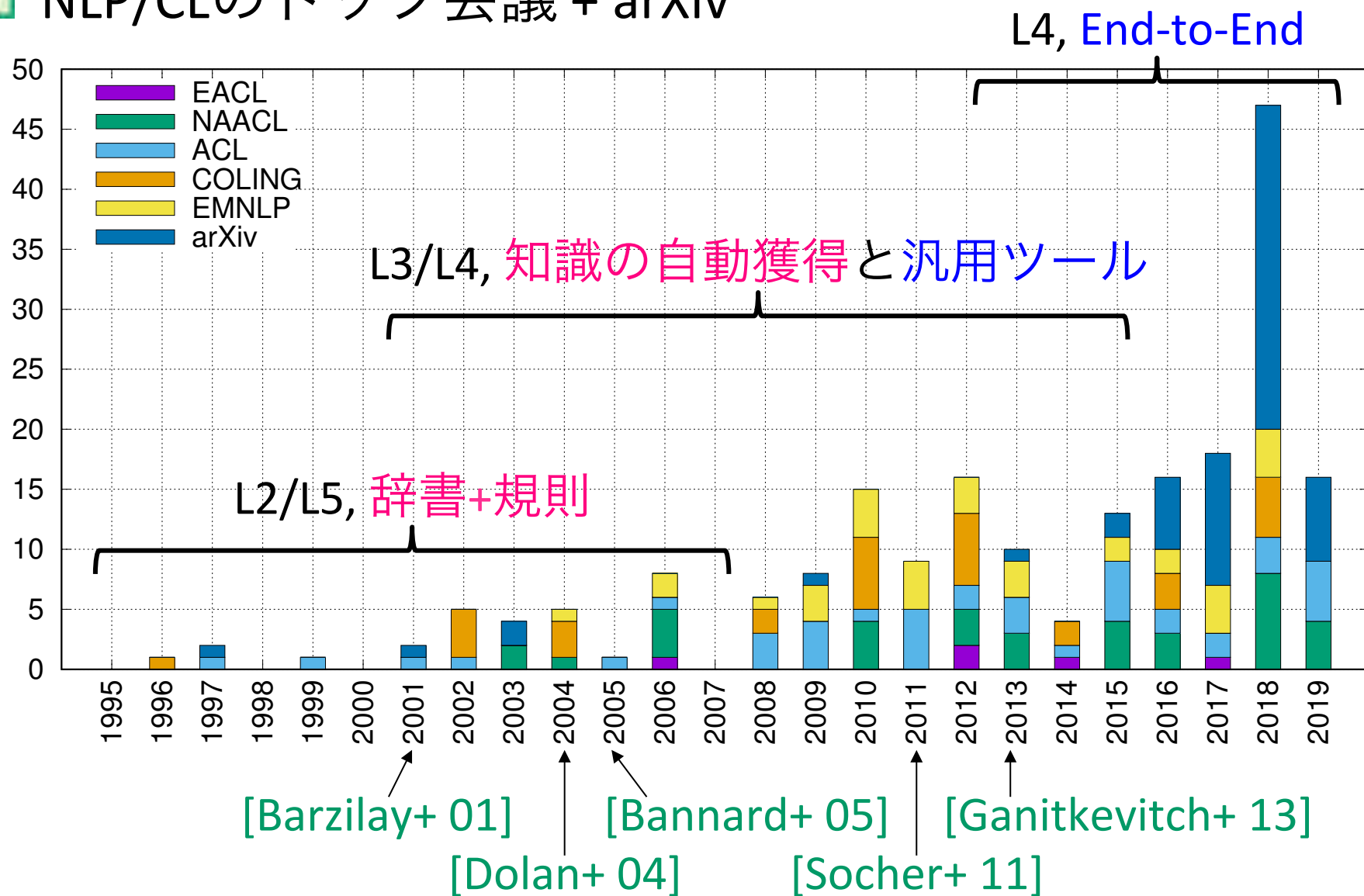
- 個々の表現は必ずしも意味的に互いに素ではない

■ 科学的興味: ことばの意味に迫るための良い例題

- 本当に迫れているかどうかはさておき

言い換えの工学的研究の歴史

NLP/CLのトップ会議 + arXiv



最近のトレンド

■ ニューラルネットワークによるモデル化

- 言い換え認識:
 - CNN [Yin+ 15][He+ 15][He+ 16][Gong+ 18]
 - Transformerを用いた事前学習 + fine-tuning [Radford+ 18][Devlin+ 18][Yang+ 19]
- 言い換え生成:
 - RNN [Prakash+ 16][Cao+ 17][Gupta+ 18][Li+ 18][Roy+ 19]
 - 制約付きデコーディング [Colin+ 18][Hu+ 2019][Kajiwara 19]
 - 出力の多様化 [Xu+ 18][Park+ 19][Kumar+ 19]

■ 評価用データセット

- Quora Question Pairs (QQP) [Iyer+ 17]
- NAACL 2019でも新たに3件
 - e.g., チャレンジセット: PAWS [Zhang+ 19]

今日紹介すること

(L5) 応用技術応用技術における活用

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

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: 具体例を集めてみる

- 網羅はできない → (L1) 現象の類型化が大事

しのぎをけずる
激しく戦う

そっくりだ
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。

どうしてこんな話をしているのか?

■ 現象や問題を定義せずに(理解せずに)解けない

- 全体像の把握
- 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?

暑すぎませんか?

窓を開けてくれませんか?

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。

- 自動化を検討する際の道標
 - 目的や応用に照らして取捨選択, 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

暑すぎませんか?

窓を開けてくれませんか?

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

聞き手のリアクション

● 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)

バルセロナの10番は先日のクラシコ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。

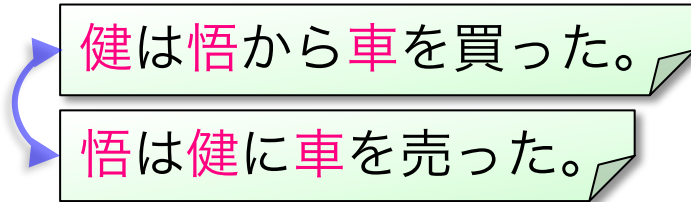
2008年～

バルセロナの視点

- 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

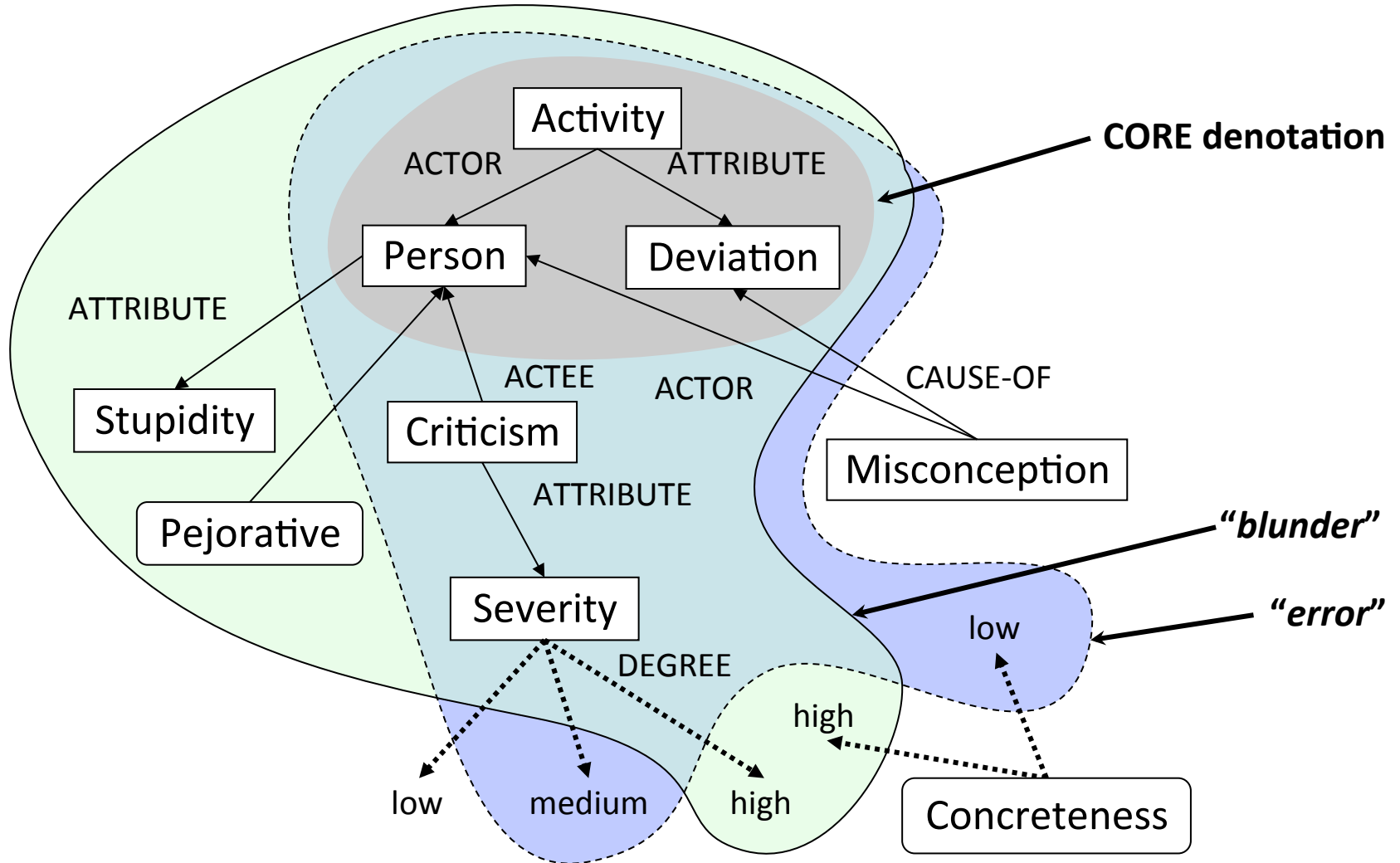
paraphrase_(v): to express in a shorter, clearer, or **different way** **what someone has said or written**

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]



違いを生じて言い換える理由

■ 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]
- 相手を騙すため

除染で取り除いた土(=汚染土)

除染土

～してください。

～していただけますと幸甚です。

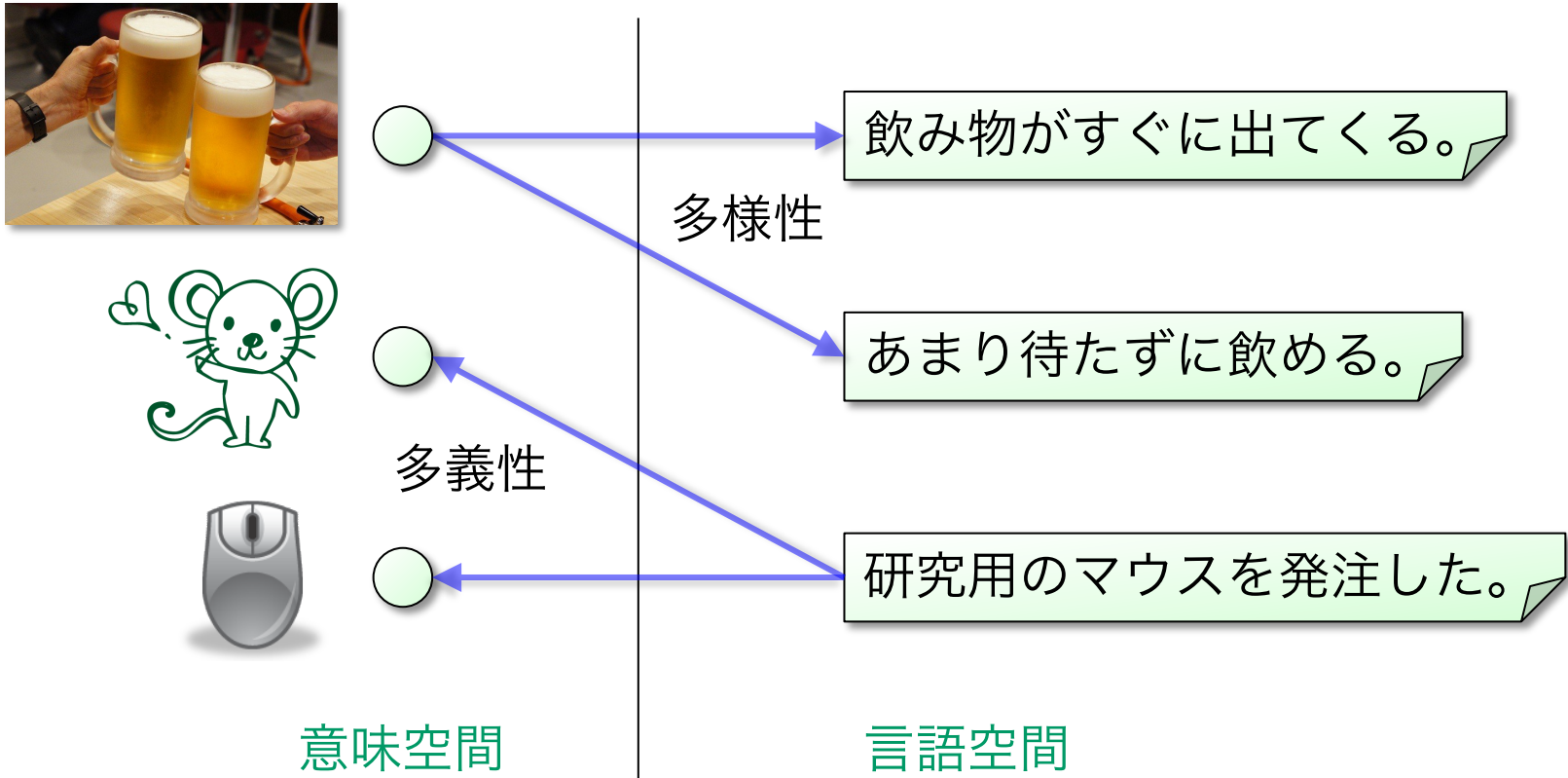
● 目的に応じた要請

- 簡単化: 難解 → 平易
 - cf. 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

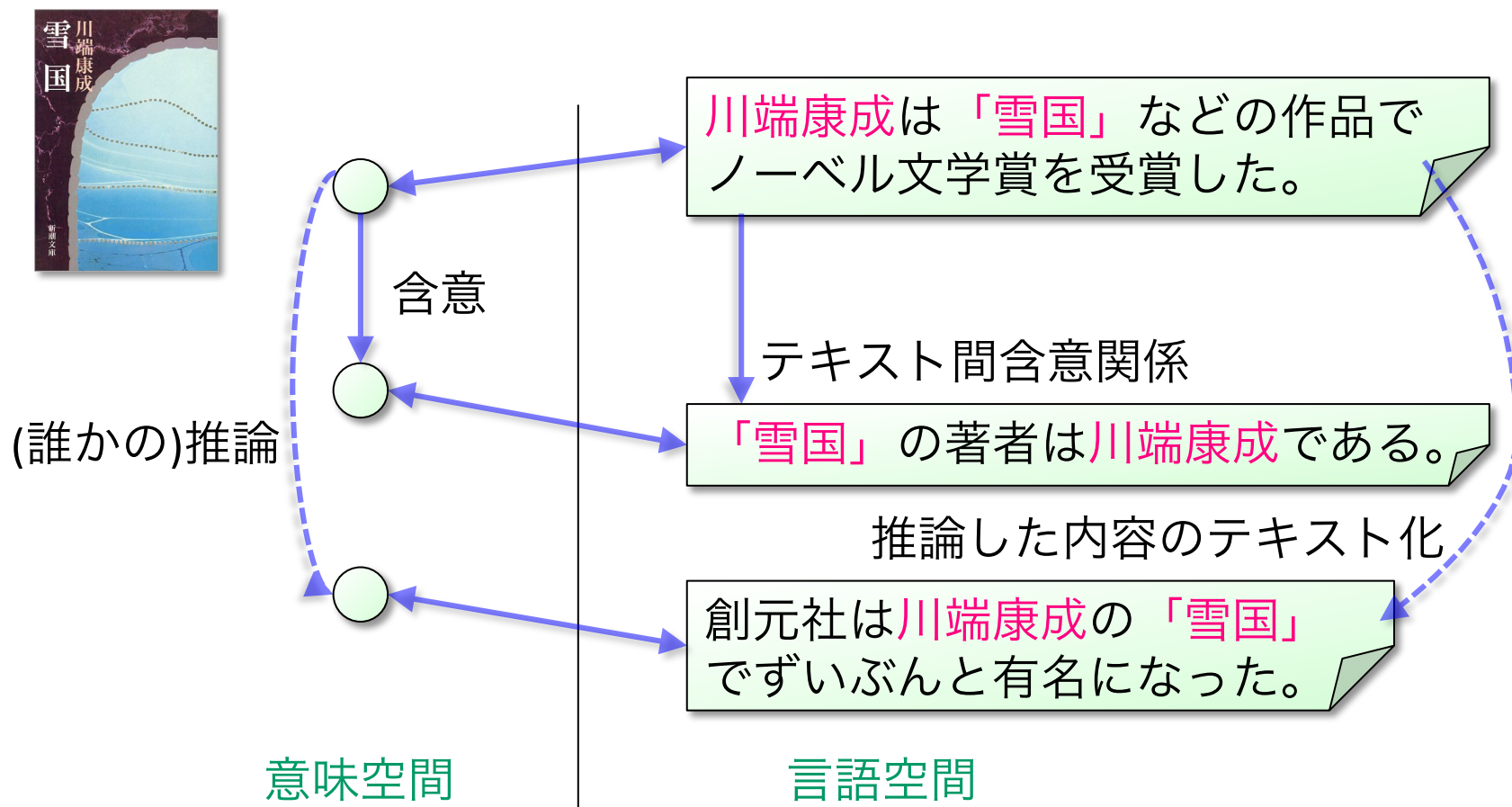
- モノ: 同じ意味内容を表す, 同言語の異なる言語表現
- コト: ある表現の言い換えを生成する行為



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)

■ $P \rightarrow Q$ (Pが成り立つならばQが成り立つ)

- $Q \rightarrow P$ も常に真とは限らない

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。

「雪国」の著者は川端康成である。 テキスト間含意関係

エベレストは世界で最も高い山である。

エルブルス山はエベレストより低い。

テキスト外の
世界知識も要参照

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

nibble

eat

具体的な方法, 上位概念

forget

know

前提

has started

started

時間的推移

- 言い換え $\equiv$ 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。

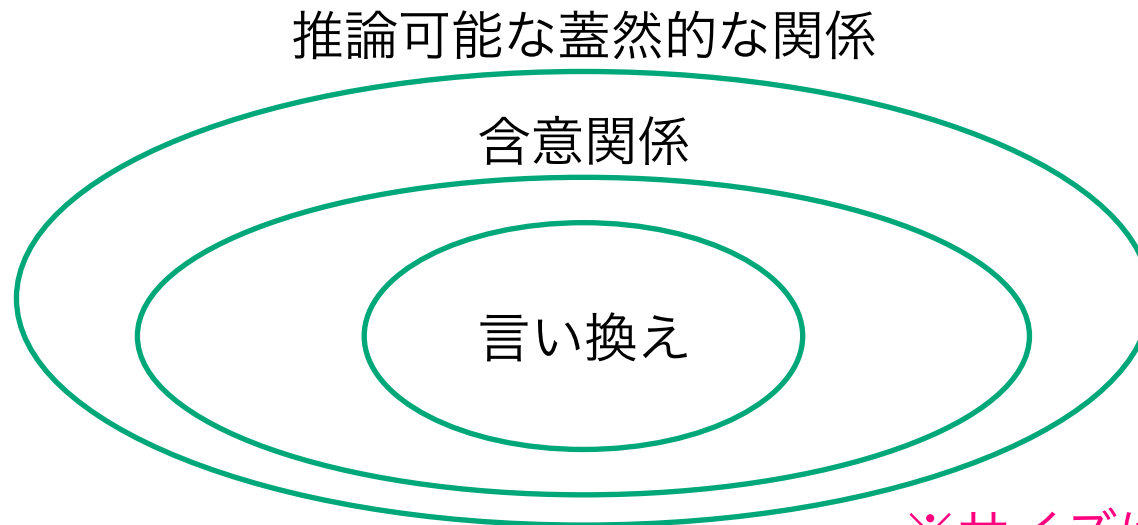
うちの子はトマトをよく食べる。

うちの子はトマトが好きだ。

このパートのまとめ

■ 言い換えとは?

- 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



※サイズは適当です

今日紹介すること

(L5) 応用技術応用技術における活用

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

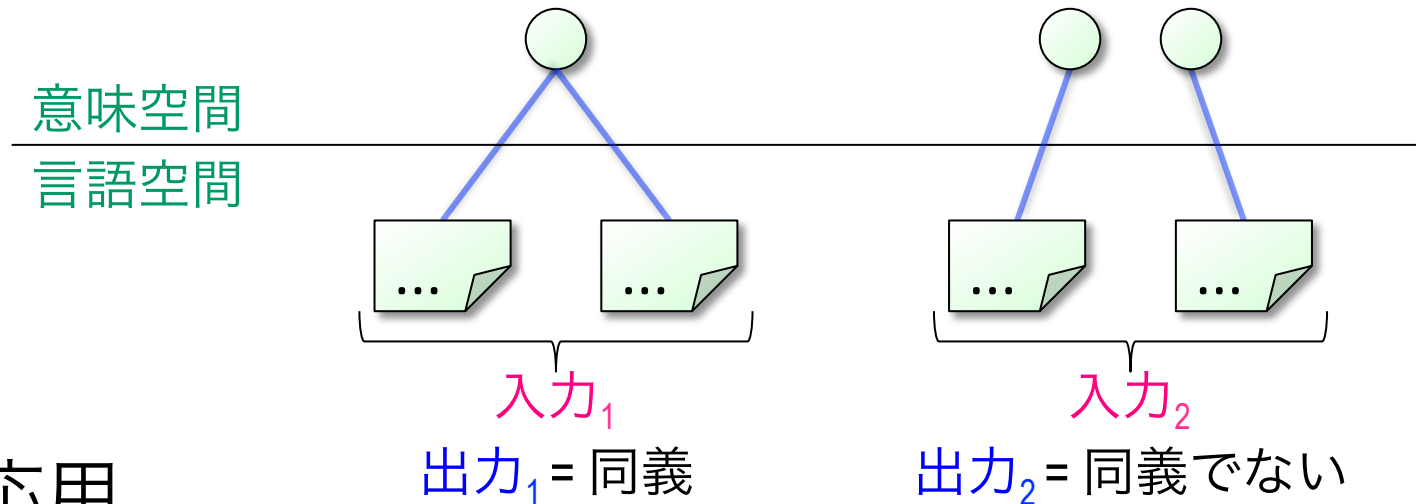
(L1) 分類体系

定義と境界, 現象の網羅・類型化, 多言語対照分析

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

- 訓練フェーズ
- 予測フェーズ

■ 性能評価

- 訓練データ同様の正解付きデータを用いる
- 2値分類問題: 再現率, 精度, それらのF値, ROCのAUC
 - 実数値予測: Pearson's r , Spearman's ρ

■ 特殊ケース: 知識の自動獲得

- 同義の文対の収集 (= テストベッドの構築)
- 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]

■ ニューラルネットワーク(NN)を用いたアプローチ

- RNNで得た文の分散表現の類似度 [Wieting+ 16]
- CNNで語レベルの対応を畳み込んで判断 [Yin+ 15]
 - 複数のCNNフィルタと複数のプーリング [He+ 15]
 - 単語類似度テンソルの畳み込み [He+ 16][Gong+ 18]
- Transformerを用いた事前学習 + fine-tuning
 - OpenAI GPT [Rdford+ 18], BERT [Devlin+ 18], XLNet [Yang+ 19]
- 生成モデルで文全体の変換尤度を推定
[Mallinson+ 17][Wieting+ 17;18]

シンボリックなアプローチ (1/3)

■ 素性

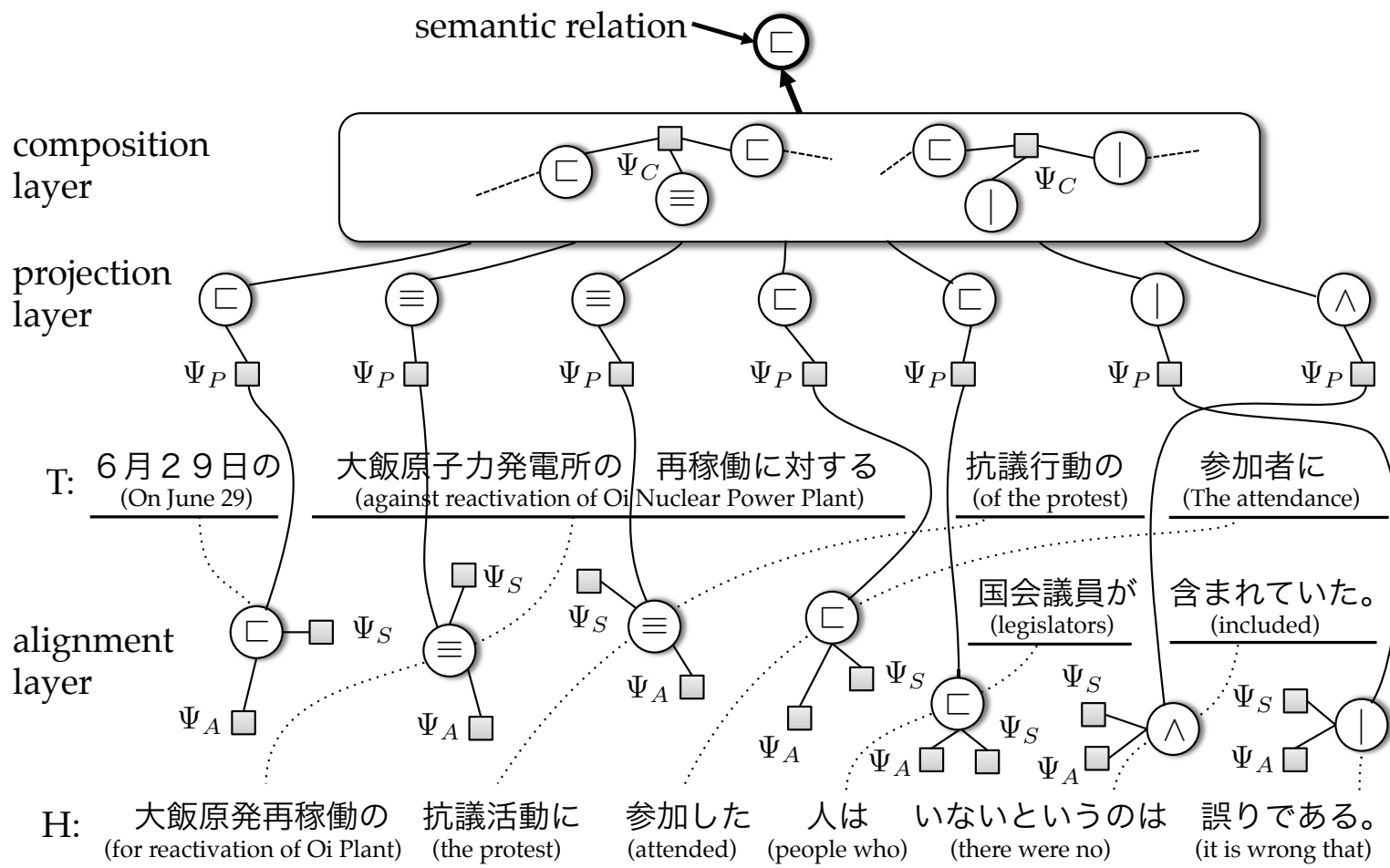
- 表層的な素性
 - MT向け自動評価尺度: BLEU等 [Finch+ 05][Madnani+ 12]
 - 表層n-gramの重複度 [Liu+ 11]
- 言語解析結果の照合
 - 統語構造 [Takahashi+ 03][Das+ 09]
 - 述語項構造 [Watanabe+ 12]
 - FrameNet [Ellsworth+ 07]
 - Abstract Meaning Representation (AMR) [Issa+ 18]

■ 学習のフレームワーク

- SVMs [Finch+ 05][Das+ 09][Issa+ 18]
 - 密な実数値素性なので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)
 - 訓練時: 非対応箇所から公理を生成

語

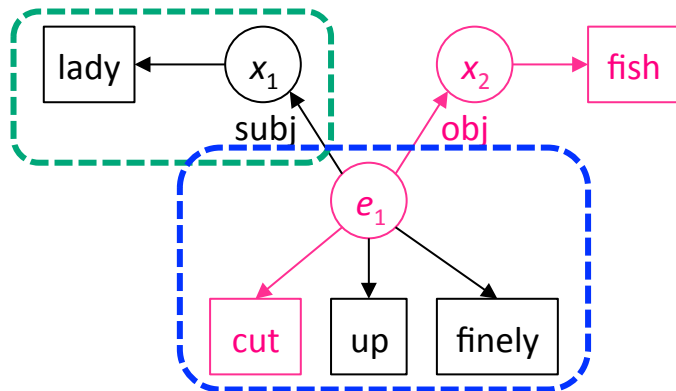
$\forall x_1 (\text{lady}(x_1) \rightarrow \text{woman}(x_1))$

句

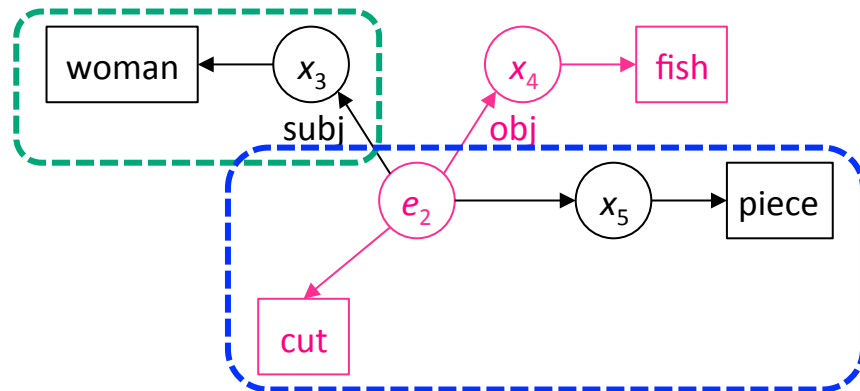
$\forall e_1 (\text{cut}(e_1) \wedge \text{up}(e_1) \wedge \text{finely}(e_1) \rightarrow \exists x_5 (\text{into}(e_1, x_5) \wedge \text{piece}(x_5)))$

- 予測時: 学習済の公理を用いて非対応箇所を単一化

A lady is cutting up some fish finely



Some fish is being cut into pieces by a woman



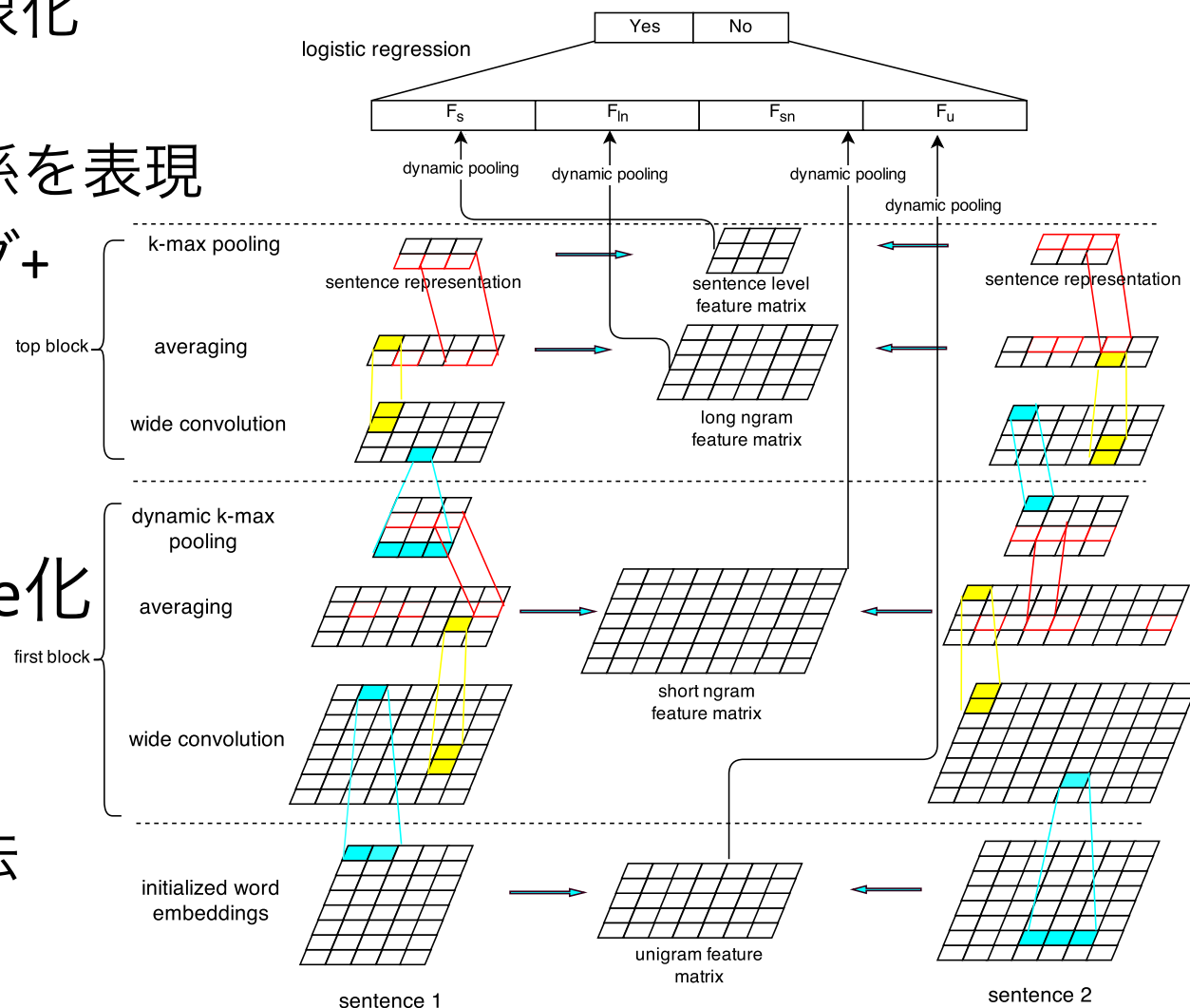
NNを用いたアプローチ (1/6)

■ 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
- 各CNNの事前学習

■ Multi-perspective化 [He+, 15]

- 類似度尺度
- プーリング方法



NNを用いたアプローチ (2/6)

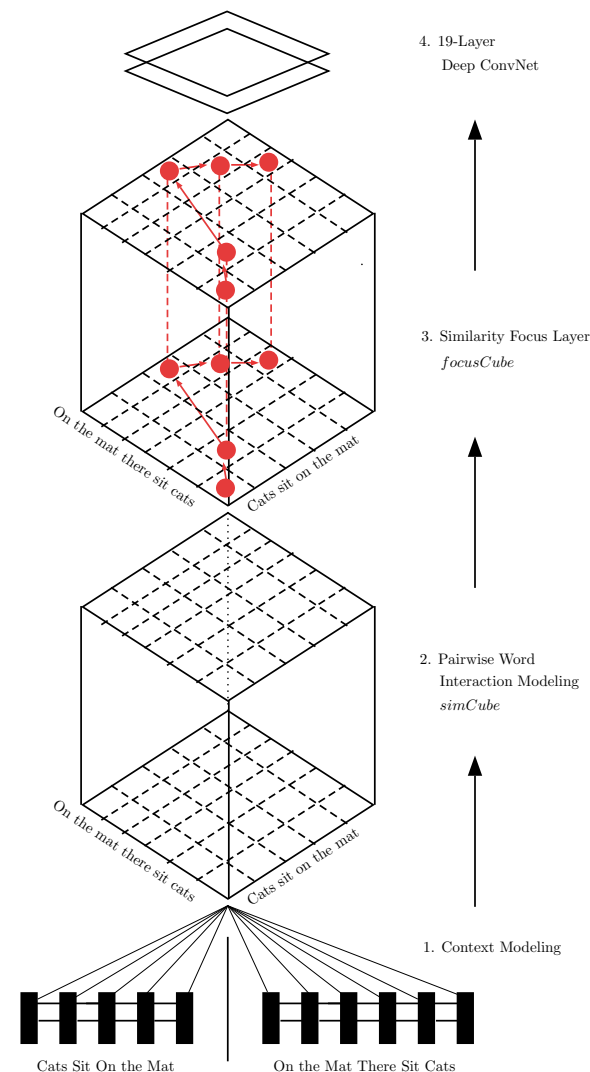
■ 文の分散表現の類似度

- 単語分散表現の平均 [Wieting+ 16]
- RNNの最終状態 [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で文内文脈を考慮 [He+ 16]
- 単語類似度行列のまとめ上げ
 - Multi-perspective (=キューブ状)
 - 各種プーリング
[Socher+ 11][Yin+ 15][He+ 16][Gong+ 18]
- 出力層: 回帰/Softmax

[He+ 16]



NNを用いたアプローチ (3/6)

Transformerを用いた事前学習 + fine-tuning

- OpenAI GPT [Radford+ 18]
 - Transformer w/ L2Rの自己注意機構
 - 連結した2文の単語相互作用を自己注意機構で捕捉

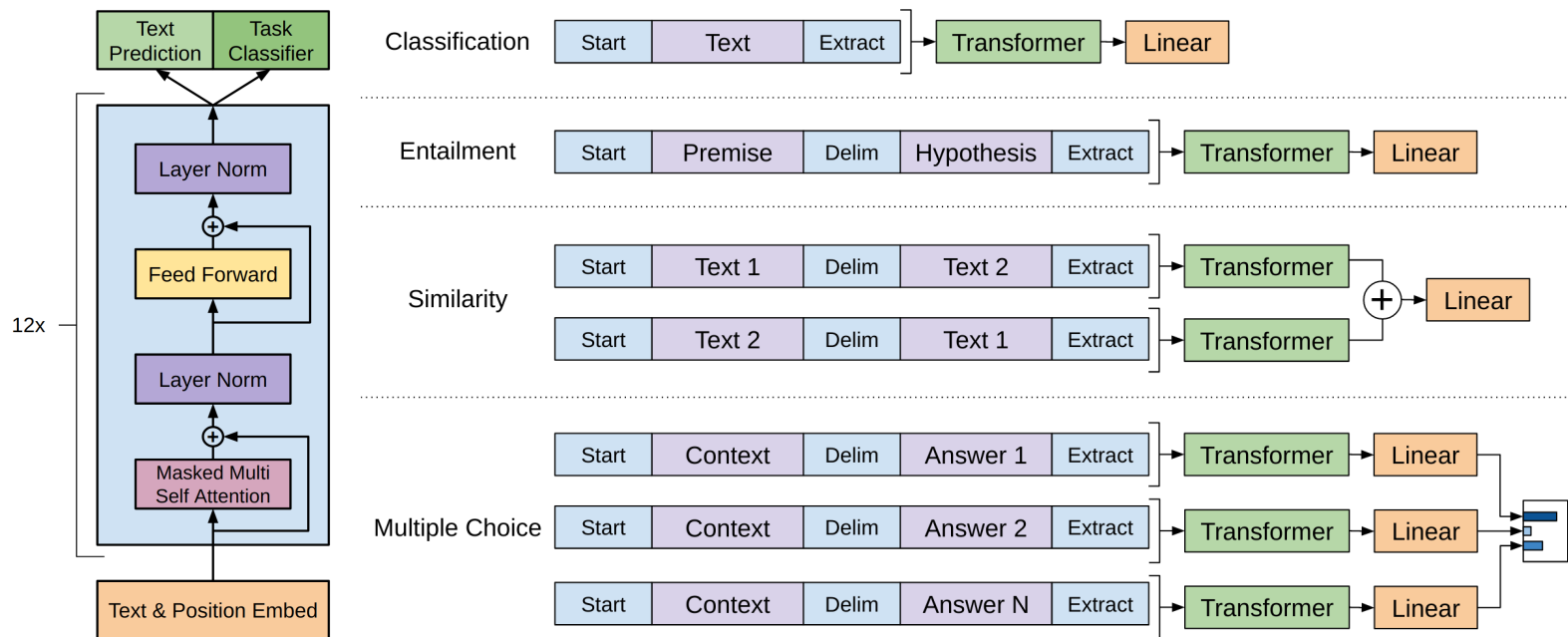


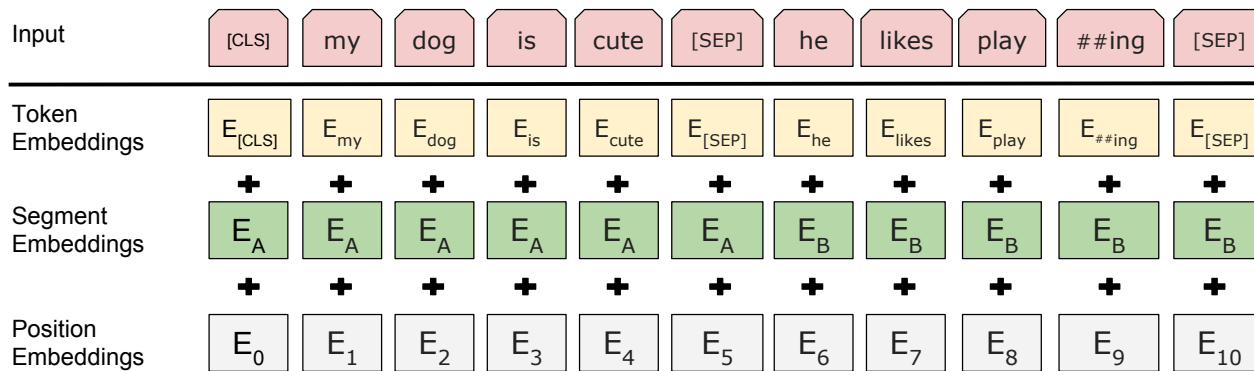
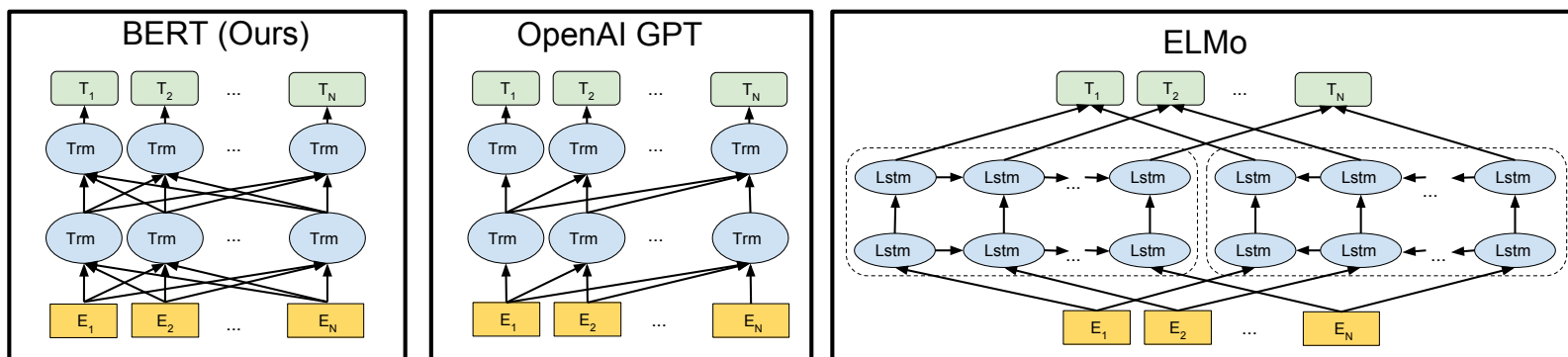
Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

NNを用いたアプローチ (4/6)

Transformerを用いた事前学習 + fine-tuning

- BERT [Devlin+ 18]

- 双方向の情報を参照するTransformer
- Masked言語モデル
- 出力層: 左端の[CLS]に対応する位置に判定結果

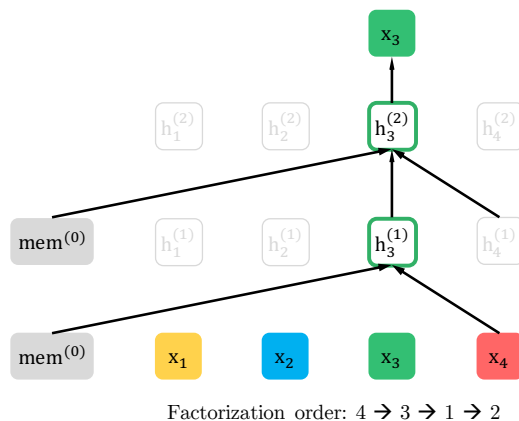
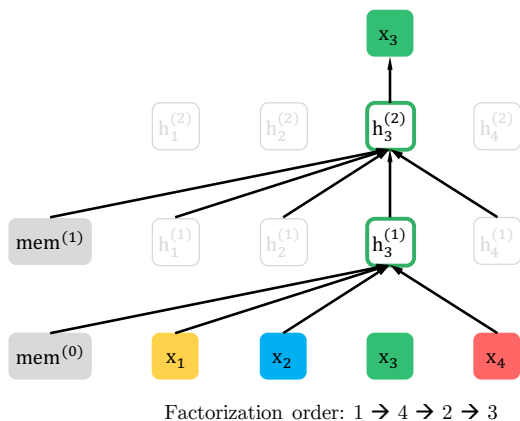
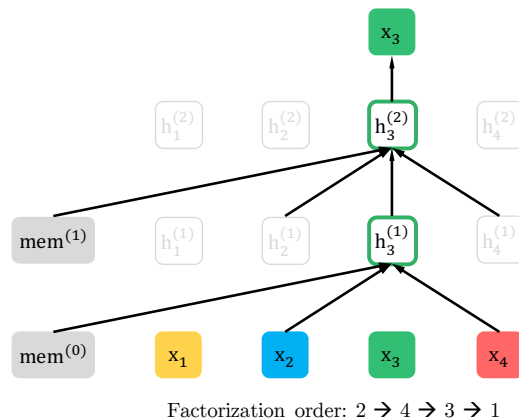
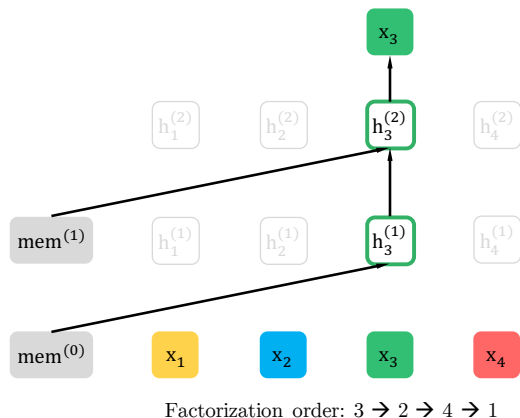


NNを用いたアプローチ (5/6)

Transformerを用いた事前学習 + fine-tuning

- XLNet [Yang+ 18]

- BERTのMasked言語モデルの疑似トークンはノイズ
- L2RとR2Lの文脈の問題をトークンの並べ替えで考慮

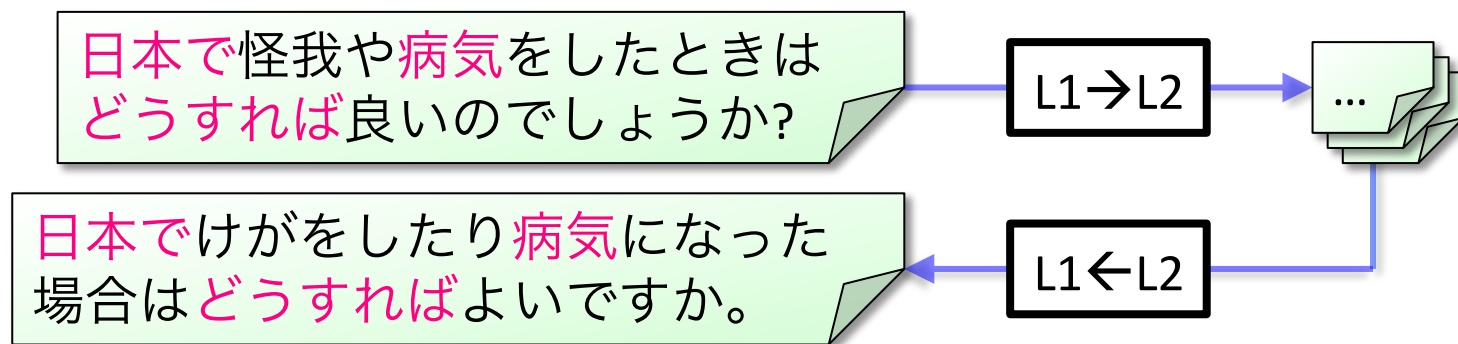


NNを用いたアプローチ (6/6)

■ 生成モデルで文全体の変換尤度を推定

- 折り返し翻訳としての尤度 [Mallinson+ 17]

- $P(s_2|s_1) = \sum_t P(s_2|t)P(t|s_1)$
 - $P(t|s_1)$: $L1 \rightarrow L2$ は n-best デコード
 - $P(s_2|t)$: $L2 \rightarrow L1$ は各 t から s_2 を forced decoding
- NMTだとPBSMTよりもforced decodingが容易



- 疑似パラレルデータからモデルを訓練 [Wieting+ 17;18]
 - L1-L2対訳データのL2をL1'に翻訳してL1'-L1データを得る

言い換え認識技術のテストベッド (1/2)

■ GLEU [Wang+ 18] に含まれているもの (英語)

- MRPC (MSRP): ニュース, 3.7k [Dolan+ 04]
- SemEval STS: 意味的類似度, 7k [Agirre+ 14][Cer+ 17]
- Quora Question Pairs (QQP): 質問文対, 364k [Iyer+ 17]

■ 言い換えの研究でよく使われているもの (英語)

- 中英MTの自動評価用複数参照訳, サイズ未確認 [Huang+ 02]
- WikiAnswers: 質問文対, 18M [Fader+ 13]
- SemEval SICK: 構成的な文間関係データ, 10k [Marelli+ 14]
- Twitter 言い換えコーパス
 - PIT-2015, 19k [Xu+ 14]
 - Twitter News URL, 52k [Lan+ 17]

言い換え認識技術のテストベッド (2/2)

■ その他, 人間の手が入ったデータ (英語)

- 小説の複数訳, 45k [Barzilay+ 01]
- MSR ビデオ記述コーパス, 85k [Chen+ 11]
- MS COCO: 画像キャプション, 330k [Lin+ 14]
- PAWS: 自動生成文の修正結果, 108k [Zhang+ 19]
- ComQA: 質問文, 4.8k クラスタ [Abujabal+ 19]
- Para-Quality: 人間の作例の分析, 6k [Yaghoub-Zadeh-Fard+ 19]

■ 日本語のデータ

- NTCIR RITE-2: 含意関係認識データ [Watanabe+ 13]
 - 開発用: 83/548, 評価用: 70/548

■ その他

- OpusParCus: 多言語字幕からの抽出物, 2k * 6言語 [Creutz 18]
- ParaNMT-50M: 人間訳とMT訳の対 [Wieting+ 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
 - 語句の重複度
 - 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]
 - 統語構造やリッチな言語資源を用いても微々たる改善
 - NNで高い精度
 - 特に大規模な単言語データでの事前訓練が有効
- “簡単すぎる”としてSuperGLEU [Wang+ 19] では除外

本当?

実は誤謬かもしれない

■ 標本バイアス: 実世界の問題の分布からの乖離

- MSRP [Dolan+ 04]
 - 編集距離が8-20の文対のみ
 - 評価用データの66.5%が正例 (常に“同義”と言えばF値79.9)
- WikiAnswers [Fader+ 13], QQP [Iyer+ 17]
 - 質問文対のみ

■ 人間の判定誤り

- “非専門家がfalse positiveを増やす” [Xu+ 14]
- 判定にかかった時間の例
 - MSRP [Dolan+ 04] は5800件を4～6人日で評価
 - PAWS [Zhang+ 19] は平均24秒/事例 → 5人で多数決
 - 事例研究向けのデータ収集 [藤田+ 06] では平均8.5分/事例
 - 現象の限定, ガイドラインあり, 不一致は議論して解消

NTCIR RITE-2 [Watanabe+ 13]

■ false positive 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 negative 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、シャープペンなどを収めた箱だ。

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。

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。

ID292
正解率19%

ID86
正解率19%

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]

(a) 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]

(b) 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 [藤田+ 06]
 - e.g., Winograd Schema Challengeの最小ペア [Levesque 11]
- 負例サンプリングのご利益 [Mikolov+ 13][Devlin+ 18]

(c) 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

(a) 評価用データの作成方法の大分類

■ 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

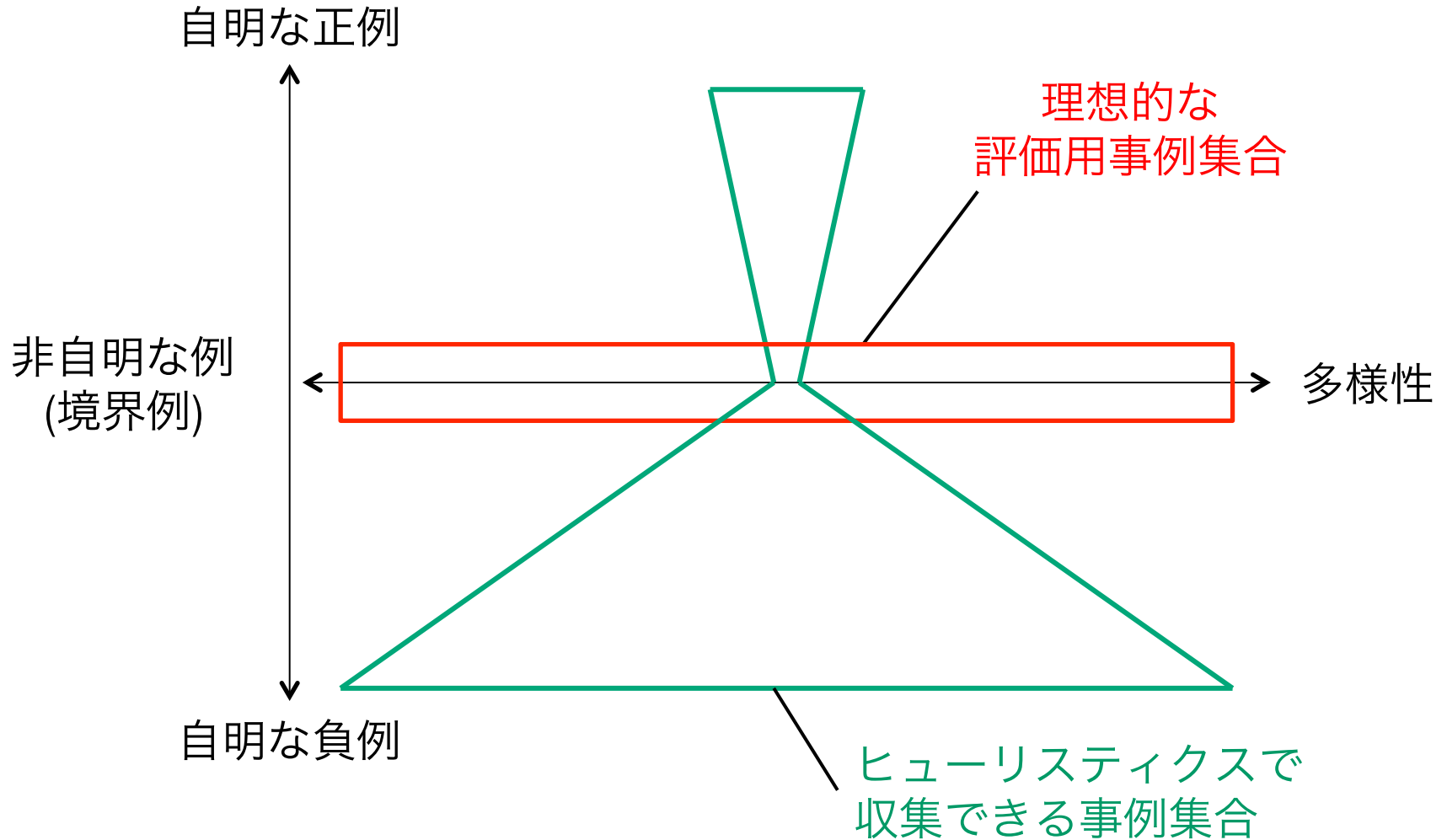
■ ある範囲の問題セットに模範解答を付与

- 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)

- cf. 人間の内省の限界 [宮田+ 17][Yaghoub-Zadeh-Fard+ 19]
- 任意のテキスト対はほぼ100%負例
- 標本バイアスの回避 vs 人的コストの低減

(a)&(b) 分布の自然さと正負例のバランス



新しい試み: 表層的な類似度からの脱却

■ Twitter上の多様な候補表現の自動収集 + 人手で判定

- PIT-2015 [Xu+ 14]

同一の語または言い換え

- 同じ時期に投稿され, アンカーを共有するツイート

- Twitter URL [Lan+ 17]

- 同じ新聞記事に言及しているツイート

- Twitterならではの特徴

- 幅広いトピック
- くだけた表現が多い

The **new Ciroc flavor** has arrived

Ciroc got a **new flavor** comin out

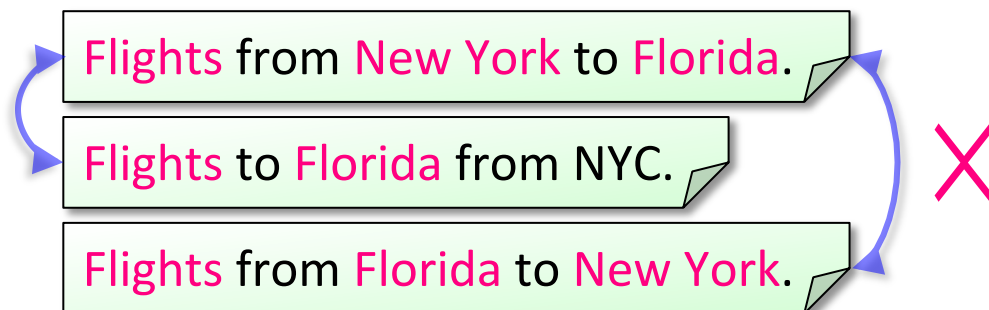
CO2 levels mark 'new era' in the world's changing climate.

CO2 levels haven't been this high for 3 to 5 million years.

新しい試み: 敵対的な事例に特化

■ 敵対的候補事例の自動生成 + 人手で判定

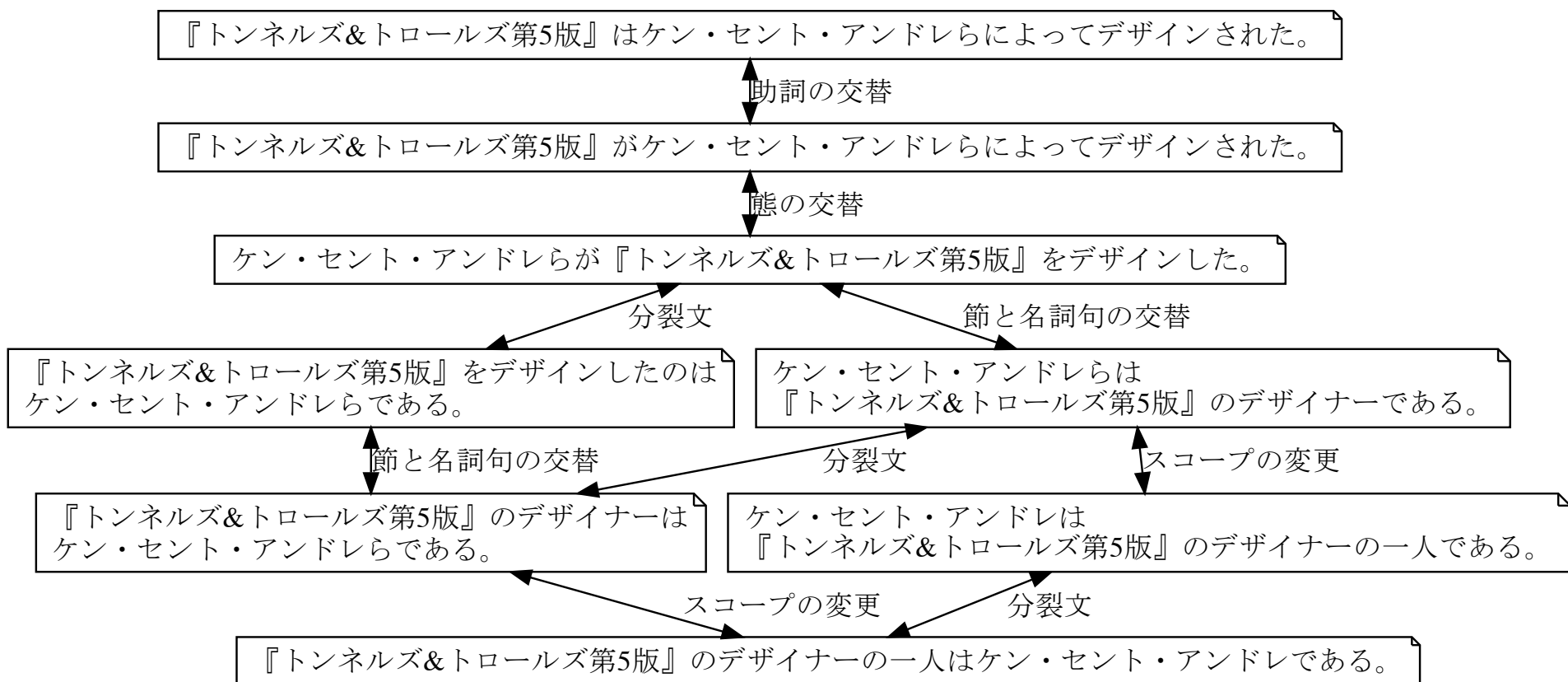
- cf. Winograd Schema Challenge [Levesque 11]
 - (真の)分離超平面近傍?
 - パターンはかなり限定される
- PAWS: 単語の並べ替え [Zhang+ 19]
 - 候補事例の自動生成
 - 品詞タグの制約つきで言語モデルによる並び替え (ほぼ負例)
 - NMTによる折り返し翻訳 (ほぼ正例) [Wieting+ 18]



(c) 事例の分解 [Kaneko+ 13][藤田+ 15]

■ どのような部分問題をどこまで解けたのかの分析用

- e.g., NTCIR RITE-2のID90-4



人間の判定誤り: 特にfalse positiveが問題

■ 何が原因?

- データ作成時のインストラクション?
- 作業者の資質? [Fujita 13][Xu+ 14][Yaghoub-Zadeh-Fard+ 19]

■ 言い換え生成の人間評価でも同じ問題が生じる

- ガイドラインの充実?
- 人数を増やして吸収?

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

このパートのまとめ

■ 言い換え認識

- 同一の言語の複数の異なる言語表現が同義か否かを判定
- 共通テストベッドの功罪
 - MSRP [Dolan+ 04] が研究を促進
 - これまでの知見
 - データは少ない → 事前学習 + fine-tuning が優勢
 - GLEU [Wang+ 18] のうち3つが関連タスクだったが SuperGLEU [Wang+ 19] では“簡単すぎる”として除外
 - 誤謬の可能性
 - 標本バイアス [Xu+ 14][Lan+ 17][Zhang+ 19]
 - 人間の判定誤り [Fujita 13][Xu+ 14][Yaghoub-Zadeh-Fard+ 19]
- どんな現象を同一視したいか?
 - 何のため?

今日紹介すること

(L5) 応用技術応用技術における活用

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

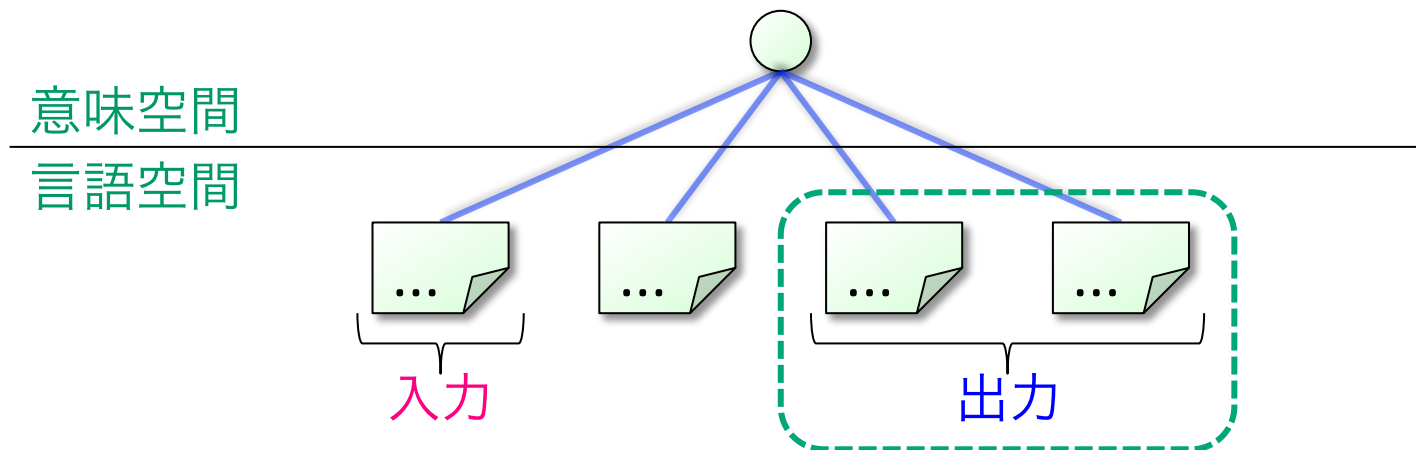
(L1) 分類体系

定義と境界, 現象の網羅・類型化, 多言語対照分析

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前編集 (e.g., 未知語→既知語)

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

(2) 仮説生成

- 必要な箇所を言い換えた候補を複数生成(過剰生成)
 - 入力の一部を書き換える
 - 入力全体を書き換える: i.e., 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

テキスト評価

■ (1)&(3) 評価基準(目的や制約)に応じた評価

- 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

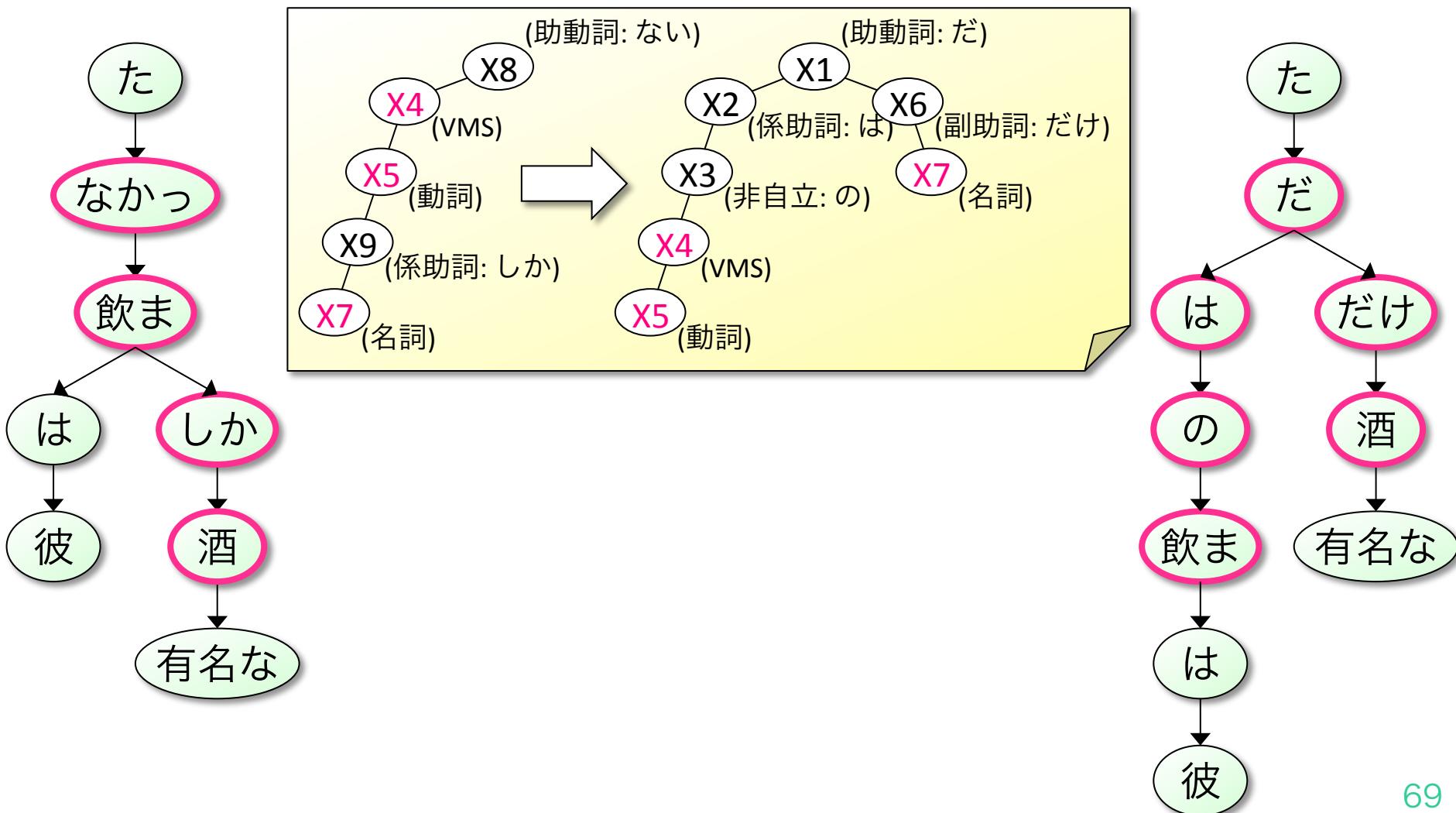
■ (3) 言い換え後処理

- 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

仮説生成手法の例 (1/5)

■ 統語構造の変換 [Lavoie+ 00][Takahashi+ 01][Zhu+ 10]

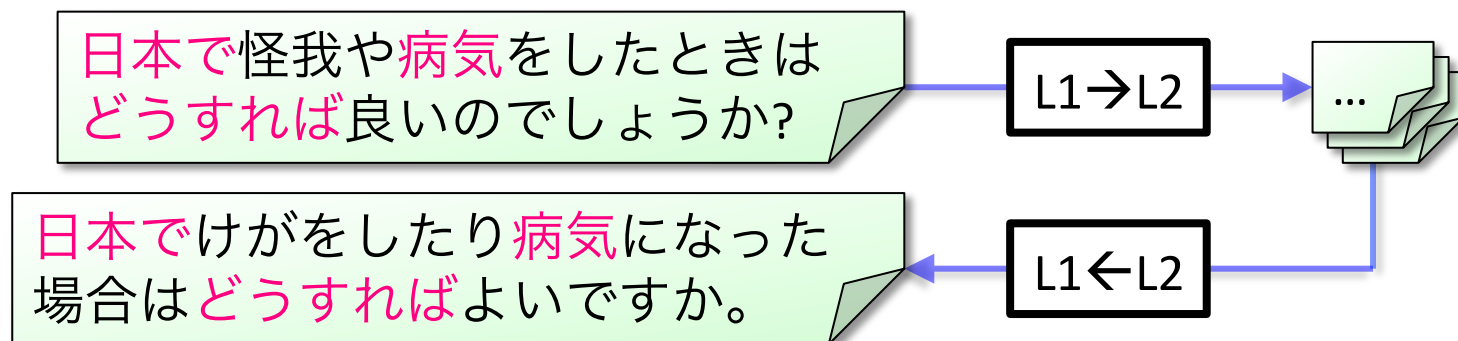
- 知識のボトルネック



仮説生成手法の例 (2/5)

■ MTによる折り返し翻訳 [Sun+ 12][Mallinson+ 17]

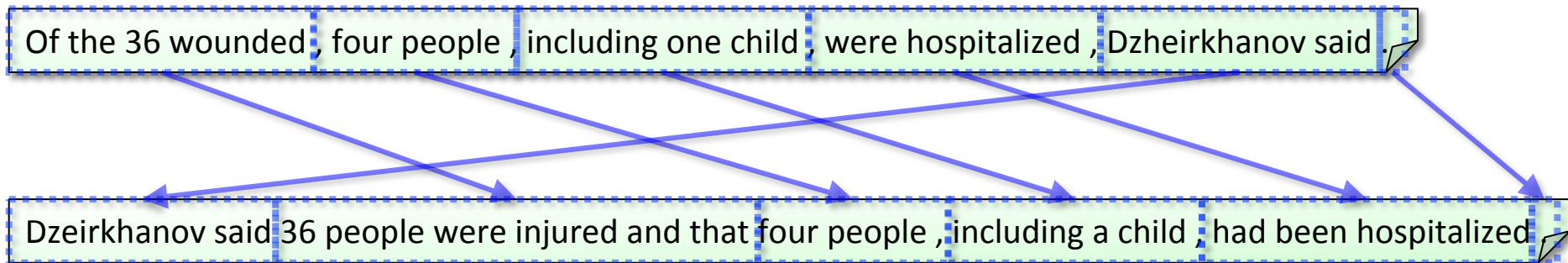
- 単言語パラレルデータは小規模にしか存在しない
→ 対訳コーパスを活用
- 人間の評価との相関: SMT < NMT
- 大規模な対訳コーパスがあれば片道でOK [Wieting+ 17;18]



仮説生成手法の例 (3/5)

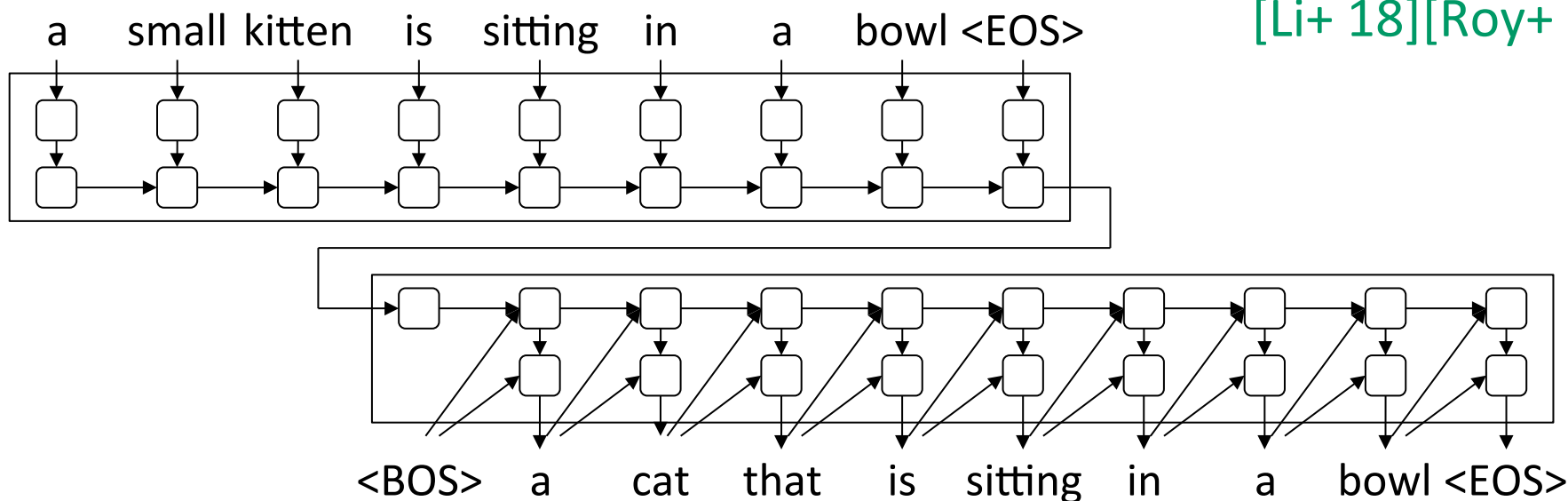
■ 単言語パラレルデータからの学習

- 句に基づく統計的MT (PBSMT) [Quirk+ 04][Wubben+ 10]

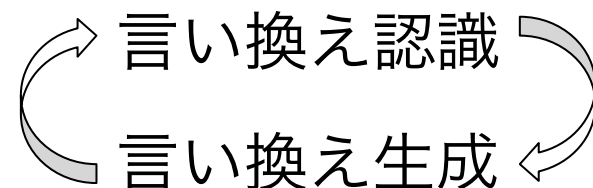


- ニューラルMT (NMT) [Prakash+ 16][Cao+ 17][Gupta+ 18]

[Li+ 18][Roy+ 19]



仮説生成手法の例 (4/5)



■ 最大マージン逆強化学習 [Li+ 18]

1. 生成器の訓練: e.g., EncDec は $P(s_2|s_1)$ を最大化
2. 評価器の訓練: i.e., 報酬関数の定義
 - 正例と負例 → 強化学習: 言い換え認識器を教師あり学習
 - 正例のみ → 逆強化学習: 生成器の出力から評価器を学習
 - Y と Y' を直接比べない (cf. 生成器の teacher forcing)
 - 類似度(X, Y') と 類似度(X, Y) の差を最大化
 - カリキュラム学習: 簡単な(類似度(X, Y)が高い)事例を重視
3. 評価器の報酬に基づいて生成器を fine-tuning

入力 (X) How far is Earth from Sun?

生成器の出力 (Y') What is the distance between Sun and Earth?

参照出力 (Y) How many miles is it from Earth to Sun?

仮説生成手法の例 (5/5)

- 変分オートエンコーダ (variational auto-encoder; VAE)
 - 最初の試み [Gupta+ 18]
 - 強化学習で多様化 [Park+ 19]
 - 教師なしニューラル言い換え生成
 - ベクトル量子化VAE [Roy+ 19]
 - 統語的制約 + マルチタスク学習
 - von Mises-Fisher-Gaussian VAE [Chen+ 19]

勉強中につき
文献のみ紹介

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
- ✗ 特定の一部のみを書き換えるという制御が困難
 - → 制約付きデコーディング

新しい試み: 制約付きデコーディング

■ 特定の表現を出力する/しないという制約

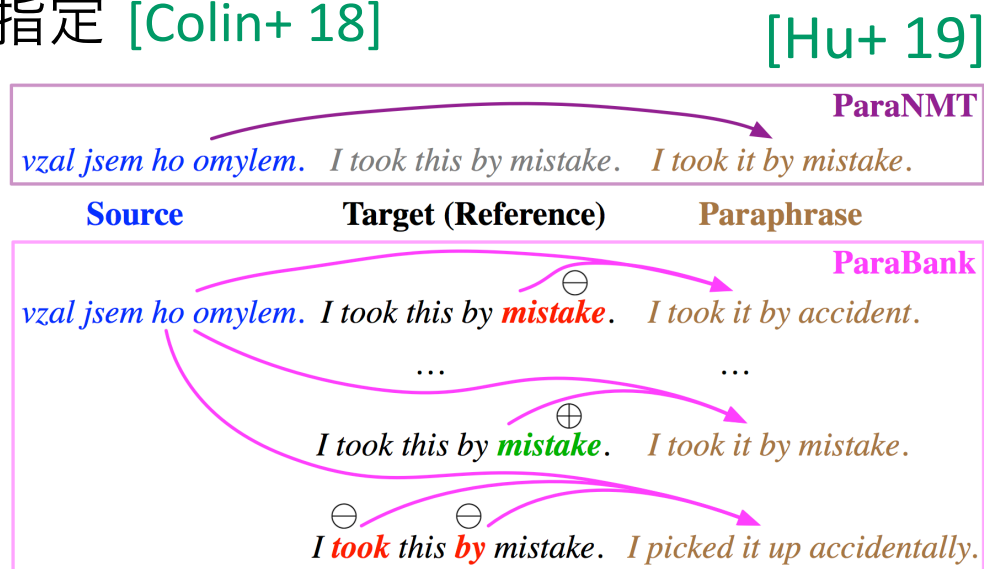
- MT用の技術 [Hokamp+ 17][Post+ 18] の応用
- 語彙的な制約: 含めたい/たくない語句
 - L2からの翻訳のガイド [Hu+ 19; ParaBank]
 - L1から直接言い換え, 出力語彙の制御 [Kajiwara 19]

● 統語的な制約:

- テンプレートを入力 [Iyyer+ 18]
- 統語的タグの加減を指定 [Colin+ 18]

- $X \rightarrow Z$ の言い換え時に
文法は別の入力Yに
似せる [Chen+ 19]

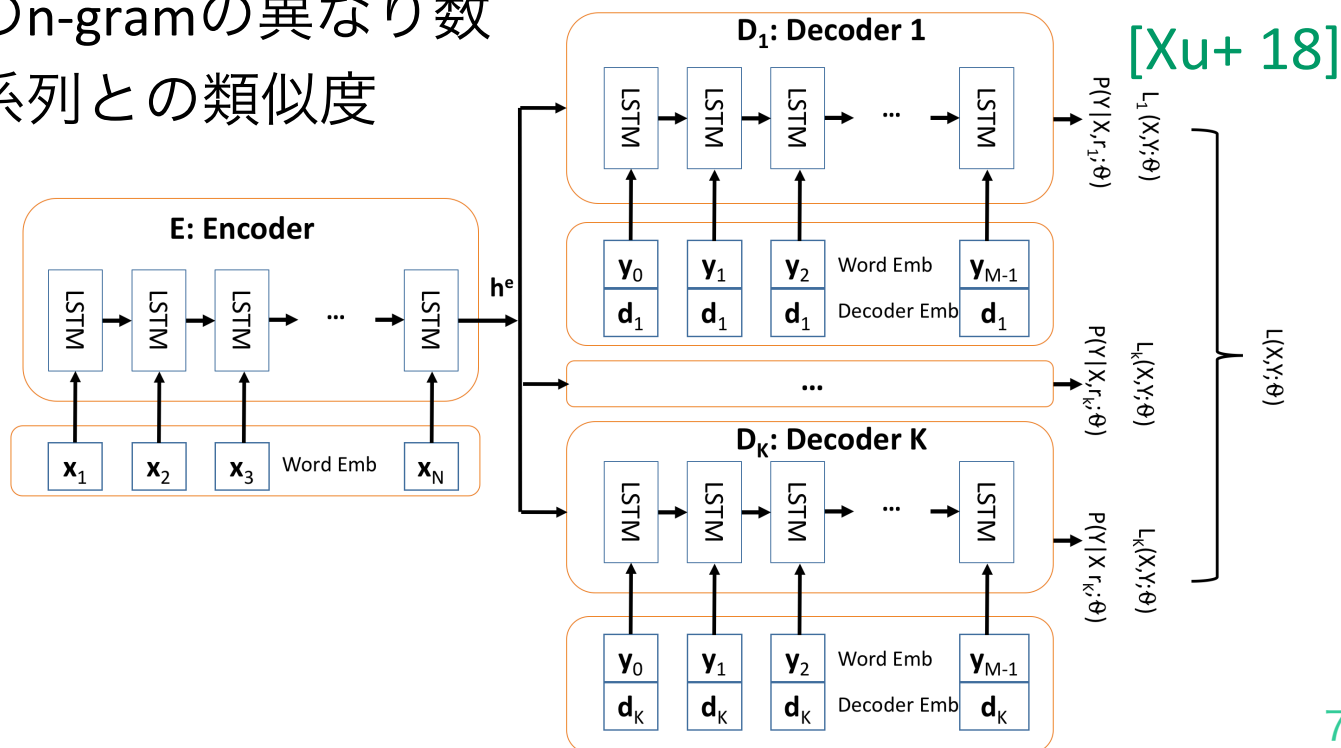
- 制約は応用側の要請
を要実装 [Inui+ 03]



新しい試み: 出力の多様化

■ テキスト生成タスクの正解は一般に複数

- なるべく多様な良い例を出力したい
- 複数のデコーダを, 互いが異なるように訓練 [Xu+ 18]
- 強化学習で揺らす [Park+ 19]
- サーチの際に劣モジュラ関数を最大化 [Kumar+ 19]
 - 候補中のn-gramの異なり数
 - 可能な系列との類似度



言い換え生成技術の性能評価

■ 出力を得るたびに人間が評価

- 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
- 決定木による人間の誘導 [Fujita 13]

■ 少数の参照出力との照合による自動評価: cf. MT

- 言い換え認識用のテストベッドの援用
- MT向け自動評価尺度の援用: BLEU, METEOR, TER
- 言い換え生成向けに提案されたもの
 - PEM: 各々のMT結果どうしの照合 [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]
 - distinct-N: 候補中のN-gramの異なり数(多様性) [Li+ 16]

言い換え生成の課題

■ 仮説生成部の改良

- 正確性と柔軟性:
 - 目的指向の制約など、色々と制御したい
 - テキスト評価に問題を先送りしない
- 多様性: やはり言語知識が必要になりそう

■ 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前編集等のデータ拡充: 非文法的でもなんとかなる

このパートのまとめ

■ 言い換え生成

- ある言語表現から，目的に応じた評価基準を満たす，入力と同じ意味，同じ言語の異なる言語表現の集合を出力
- NNによる仮説生成: MT向け技術の援用
 - 技術的進展: (逆)強化学習, VAE, 教師なし
 - 新しい試み: 制約付きデコーディング, 出力の多様化
- 課題
 - 正確性, 頑健性, 多様性の両立
 - 効率的かつ再現性のある評価手法の確立
- どんな種類の言い換えを生成したいか?
 - 何のため?

言い換えに関する研究の全体像

■ 今日紹介したこと

(L5) 応用技術応用技術における活用

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

研究課題: 工学的な観点から

■ (L5) 応用から考える

- (L1) 解決したい現象を定義する → 言い換え? その他?
 - 言い換え分類体系はここに寄与
 - 自動化できたときのご利益を見積もる
- (L4)&(L5) 評価基準を定義する
 - 言い換え認識? 言い換え生成? その他?
 - (L3) 評価用データセットも構築する

■ (L4) 計算機構の高度化

- (L4) アルゴリズム: 表現の多義性や文脈の考慮
- (L4) 評価手法: 効率的な手法の確立(特に言い換え生成)
- (L3) 知識開発: 知識の形式のデザイン, 資源の集約
 - (L2) 言い換えの種類とセットで

(L5) 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 (e.g., “スラっとした” vs. “ガリガリの”)
 - 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

(L3) 知識の整備

■ 言い換えの「自動処理」に知識は必要か？

- End-to-Endのアプローチでは無視されがち
- 不要ということが示されたわけではない
 - 陽に記述しないと扱えない言語現象の存在
 - 知識の使いみち
 - End-to-Endのモデルの外側
 - 訓練データへの注入

研究課題: 言語学的な観点から

■ (L1) 包括的な分類体系の構築

- 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 (cf. [Levin 93])

