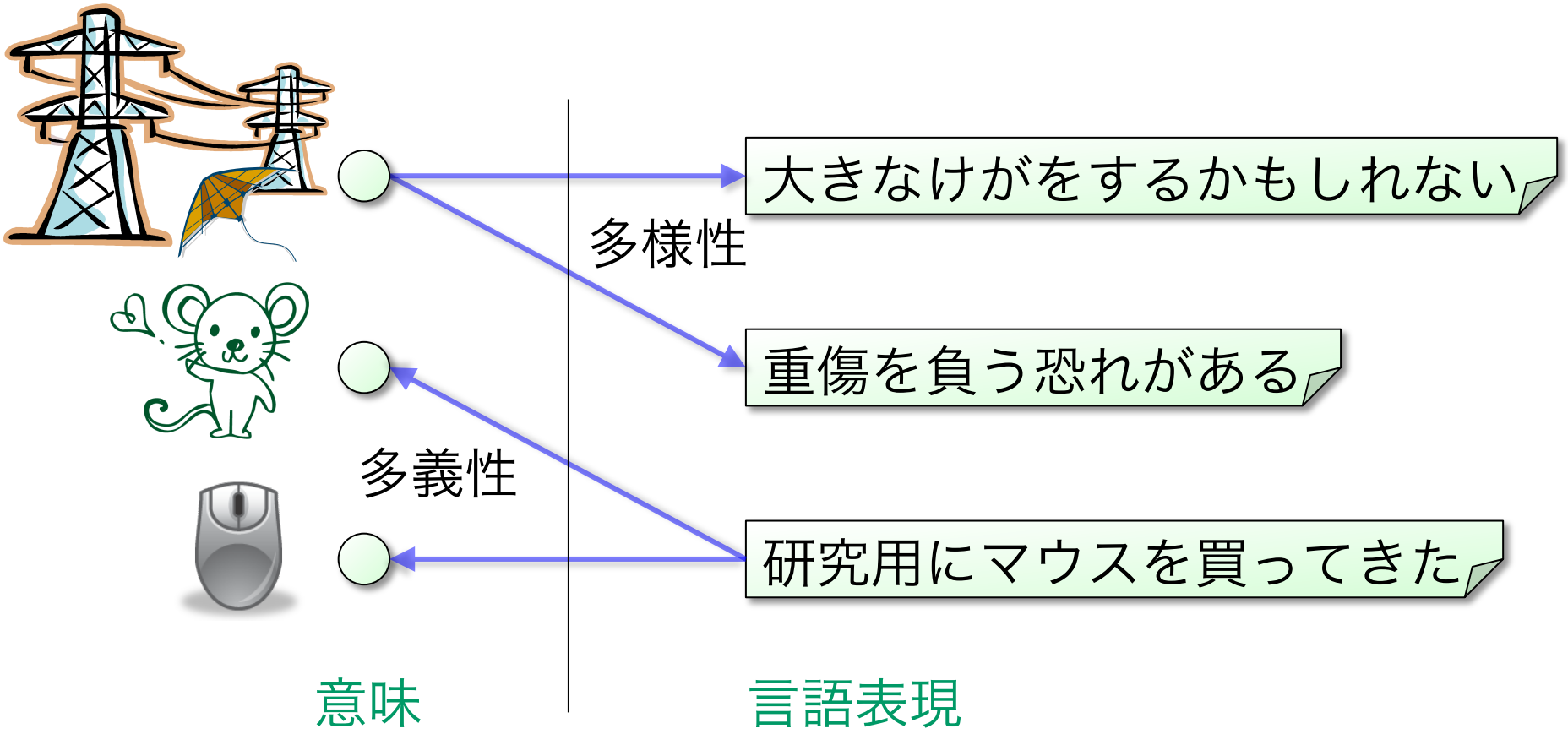
Project Next NLP 言い換えタスク: エラー分析用の primitiveな事例集の構築

藤田 篤 (NICT)　柴田 知秀 (京大)　松吉 俊 (山梨大)
渡邉 陽太郎 (NEC)　梶原 智之 (長岡技科大)

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



言い換えの(一般的な)サブタスク

- 言い換え生成
  - [入力] 言語表現 → [出力] 入力と同義の言語表現の集合
    - 目的に適った(あるいはあらゆる)言い換えを生成
  - 応用先: 簡単化，圧縮，機械翻訳の前処理，折句生成
- 言い換え認識
  - [入力] 2つの異なる言語表現 → [出力] ラベル ∈ {＝, ≠}
    - 同じ意味を表すか否か (あるいはその程度)
  - 応用先: 情報検索，質問応答，複数文書要約，剽窃検出
- 言い換え知識獲得 (言い換え認識の特殊版)
  - [入力] コーパス等の言語資源
    → [出力] 同義表現(語，句，文)の集合の集合
  - 上の2タスクの根幹，あるいは応用に直接貢献

日本語タスクにおけるエラー分析

- 言い換え認識タスクに着目
  - 言い換え生成の評価は難しい
    - 言い換え認識よりも複雑 (タスク指向，多数の解)
    - 現在の技術で生成できる言い換えは限定的
    - 評価データも存在しない

	技術	評価・分析用データ
言い換え生成	KURA [Takahashi+, 01], SMTの援用による単語列の置換と並べ替え [Quirk+, 04][Wubben+, 10]	言い換えのあれこれ [藤田, 12]
言い換え認識	Tree Kernel [Collins+, 01], 一般の分類器/回帰	NTCIR-10 RITE2 の一部 [Watanabe+, 13]
言い換え知識獲得	色々	存在しない

分析済 [藤田+, 03]

対象が限定的

様々な手法

頻度は未考慮

評価データとシステムの出力

NTCIR-10 RITE2 のデータに基づく分析

Recognizing **I**nference in **TE**xt

RITE2の多クラス分類問題

- 言い換えを含む4クラスの分類問題
  - **B**idirectional entailment (= 言い換え)

    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。 ID=26
  - **F**orward entailment

    肥満は悪性腫瘍、脳卒中、心臓病の疾患になるリスクを上げる。

    肥満は数多くの疾患のリスクファクターとなる。 ID=5
  - **C**ontradiction

    東ハトは，クイズ大作戦の正式なスポンサーから外れている。

    クイズ大作戦は，東ハトが一社提供している。 ID=64
  - **I**ndependence

    陽極泥とは電解精錬を行った際、得られる不純物である。

    電解精錬では、陽極の下に陽極泥が沈殿する。 ID=52

参加システムの性能

- 言い換えクラス(B)の認識性能はまだ低い
  - 最高性能はF値69.3: 文字の重複に基づく手法 [Hattori+, 13]
    - cf. Microsoft Research Paraphrase Corpus [Dolan+, 04]
      を用いた英語タスクではF値84.1 [Madnani+, 12]
  - 他のシステムで用いられた言語資源
    - 要素技術: アラインメント，述語項構造解析，機械翻訳(!)
    - 語彙資源: WordNet，反義語，含意等の辞書，Webコーパス
  - エラー分析が必要
    - どんな問題が解けている/いない?
    - 言語資源は活用できている? そもそも足りている?

解ける事例・難しい事例

- 事例データは公開済
  - NTCIR-11 RITE-VALオーガナイザ (松吉，柴田，渡邉) による
    - *http://warehouse.ntcir.nii.ac.jp/openaccess/rite/10RITE-Japanese-wiki.html*
  - 開発用: 548事例 (言い換えは83事例)
  - 評価用: 548事例 (言い換えは70事例)
- 参加システムの事例ごとの正答/誤答状況の分析
  - NTCIR-10 RITE2オーガナイザ (渡邉，柴田) の監督下
    - 評価用データに対する参加21システムの出力は非公開
  - B/F/C/I の分布 → 相対的な容易さ/難しさ
    - B (言い換え): 正解
    - F (一方向): 惜しい
    - C/I (上記以外): 不正解

比較的よく解けている事例

- ID=242: B(17), F(3), C/I(1)
  - 格要素と従属節の順序の変更

太平洋戦争の敗戦に伴い、陸軍幼年学校は廃止され、解散した。

陸軍幼年学校は、太平洋戦争の敗戦に伴い廃止され、解散した。

- ID=186: B(15), F(4), C/I(2)
  - 対義語の入れ替え

公社債投資信託の対義語は株式投資信託である。

株式投資信託の対義語は公社債投資信託である。

正解率がやや低い事例

- ID=199: B(12), F(2), C/I(6), 未回答(1)
  - 対義語の入れ替え + 同格の括弧表現

紅色組合の対義語は御用組合（黄色組合）である。

御用組合は、俗に黄色組合とも言われ、対義語は紅色組合である。

- ID=330: B(10), F(2), C/I(9)
  - 態の交替 + 名詞句/名詞

未成年者喫煙禁止法によって未成年の喫煙は禁止されている。

未成年者喫煙禁止法は，20歳未満の者の喫煙を禁止している。

ほとんど解けていない事例

- ID=292: B(4), F(10), C/I(7)
  - 並列要素の並び替え + 名詞/名詞 + 助動詞 + …

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。

- ID=86: B(4), F(3), C/I(13)

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。

- ID=26: B(2), F(7), C/I(12)

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。

現時点でのエラー分析は建設的ではない!?

- 複数の部分問題を含む複雑な事例
  - どのような部分問題がどこまで解けているかが分からない
    - 単純な問題: 解ける/解けない理由を説明できる
    - 複雑な問題: なんだかうまく解けた/なんだか難しい
  - 個々の部分問題は必ずしも相互依存していない
  - そもそもどんな部分問題が含まれているかが自明ではない

- ブラックボックスシステム
  - 多様なシステムはあっても中身は見られない
    - 使用される語彙知識やその組み合わせ方はおそらく複雑
  - 個々の部分問題を解くのに適切なアプローチか?

エラー分析に向けての戦略

エラー分析に向けての戦略

- 目標: 何がどこまで解けているのかを
  (より平易に)評価できるようにする
  - Step 1. primitiveな言い換え事例集の構築
    - とことん分解した事例の収集(作例)・蓄積

現時点での進捗

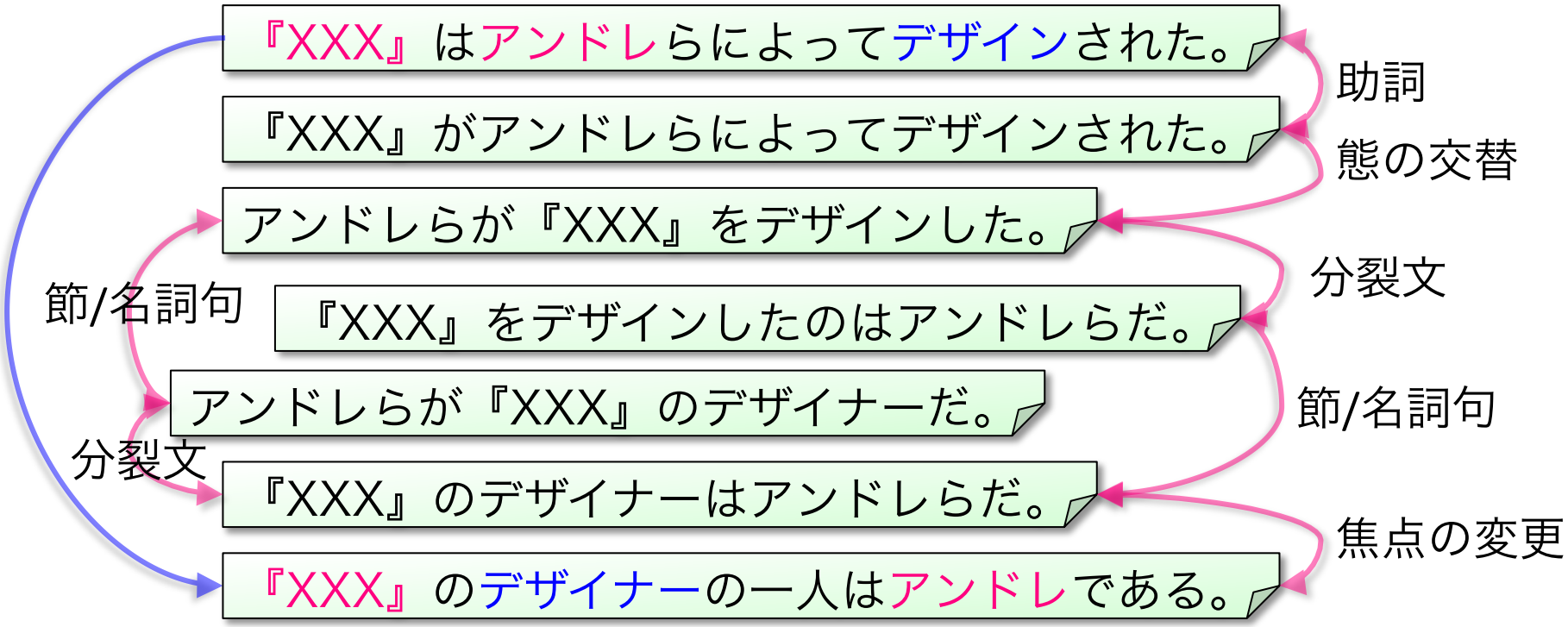
  - Step 2. 必要な知識・機能を列挙 [Sammons+, 10]
    - (人間ならば)どのような知識・処理に基づいて判断できるか
      - cf. 人間のプロセス ≠ 理想的なシステム
    - 知識・処理の具体的なインスタンスを列挙
  - Step 3. 現在の技術動向の達成度の評価と課題の提言
    - a. 既存の語彙資源の評価
      - 既存のリソースに，必要な語彙知識が含まれているか否か
    - b. 各種手法のエラー分析
      - 特定の手法で，どのような問題が解けているか/いないか

進捗状況報告

- Step 1. primitiveな言い換え事例集の構築
  - RITE2のユニットテストデータをさらに分解
    - 本活動と同じ意図で作成されたデータだが，まだ複雑
    - 含意か否かの分類 (言い換えではない)
    - 61事例由来の241件，分析対象は163件
  - 相互依存性がある場合は分解しない
  - 分解後の(primitiveな)言い換えの分類
    - Ref.『言い換えのあれこれ』の大分類8種類，小分類約40種類
      - *http://paraphrasing.org/paraphrase.html*
    - まずは各作業者が適当に名付けて，あとで調整

事例の分解例

- ID=90-4, Type=synonymy:phrase



事例の分解・分類の結果 (9月3日10:30現在)

- 163件中118件 → 262件のprimitiveな事例
  - 内訳
    - 45件: 分解なしで非言い換え (e.g., 文の一部の抽出)
    - 58件: 分解なしで言い換え
    - 60件: 204件のprimitiveな事例
  - 49種類
    - 元々のデータ自体が人工的であるため偏りあり
      - 助詞の変更(32)
      - 自明要素の明示/暗示(24)
        - ✓ e.g., ゼロ代名詞，主題の主辞のコピュラ
      - 主題の交替(19)
      - 格要素の語順の変更(13)
      - 態の交替(11)
    - TODO: 『言い換えのあれこれ』との比較

今後の活動予定 (年度末までの目標)

- Step 2. 知識のインスタンスの列挙
  - どのような知識・処理に基づいて判断できるか
    - cf. 人間のプロセス ≠ 理想的なシステム
  - 知識・処理の具体的なインスタンスを列挙
- Step 3a. 既存の語彙資源の評価
  - 必要な語彙知識のインスタンスが含まれているか否か
    - 言語知識，文書内文脈，世界知識の境界の整理
- 各種資源の公開 @ *http://paraphrasing.org/nextnlp/*
  - 分解済み事例データ
  - 言い換えの種類のリスト
  - 各言い換えに必要な語彙資源のリスト

今後の活動予定 (中期的な目標)

- Step 3b. 各種手法のエラー分析
  - Free Paraphrase Recognition Task の実施
    - 多様で state-of-the-art なシステムの参加を募る
    - cf. Kyoto Free Translation Task [Neubig, 11]
- 評価用データの課題
  - 境界をうまくとらえるような負例も必要 [Zaenen+, 05]
    - 言い換えクラスごとの半自動事例生成/収集 [Fujita+, 05]
    - 49種類全部!?
  - 分布の自然さの担保 (type vs. token)
    - RITE2由来の偏りの排除

Task 2. 言い換え生成タスクの正解データ

- ある範囲のテキストをとことん言い換えてみよう
  - 今回作成したデータも援用可能
    - 分解して得られた言い換えのあらゆる対
      - ただし，偏りがある点には注意
  - TODO:
    - 対象データ・規模 (BCCWJの一部，他のタスクとの共有)
    - 作成指針 (請負作業発注の際の仕様書)
    - 支援・事例蓄積ツール