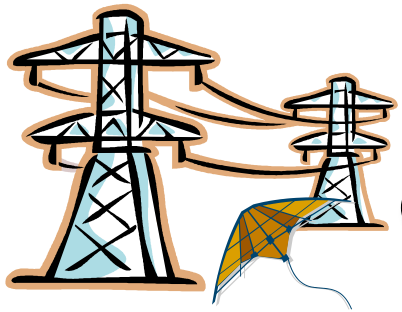
言い換え認識技術の評価に適した言い換えコーパスの構築指針

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



多様性

多義性

大きなけがをするかもしれない

重傷を負う恐れがある

研究用にマウスを買ってきた

意味空間

言語空間

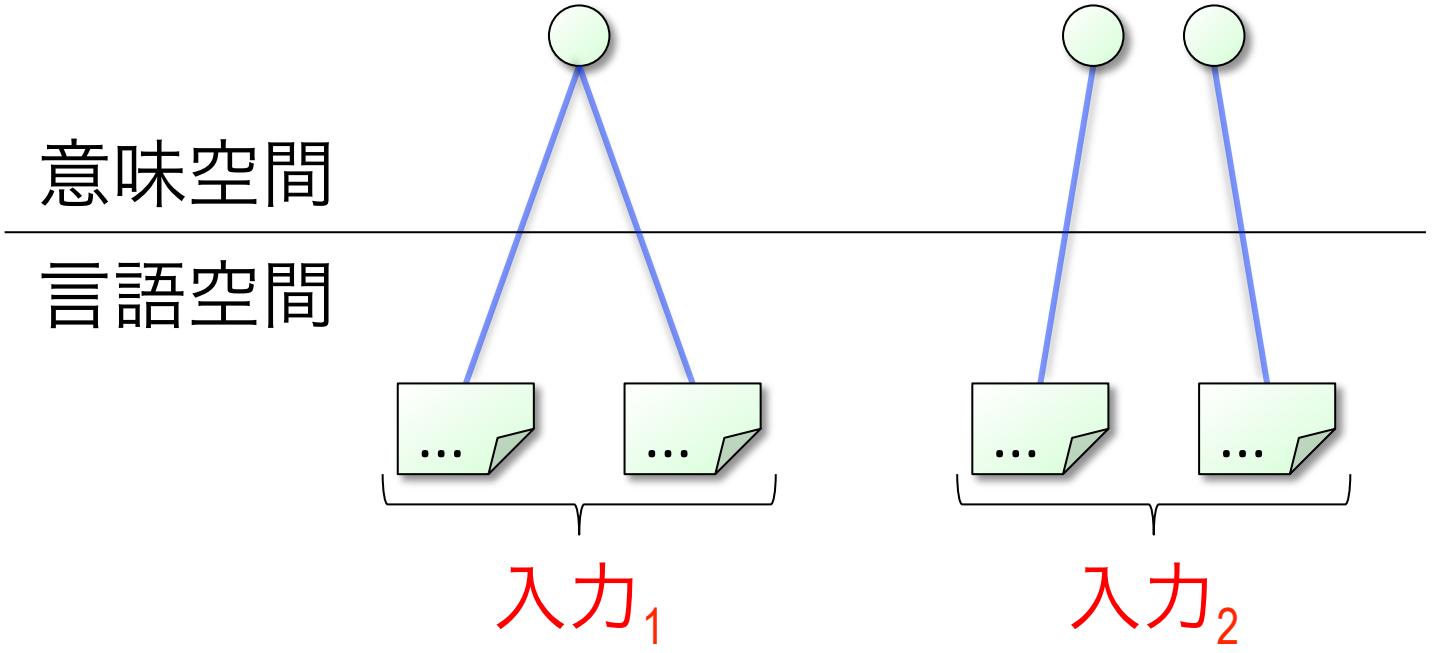
言い換え技術に関する研究動向

- とある分類
  - *http://paraphrasing.org/bib-cat.html*
  - 大分類12，小分類50，マルチラベル
  - 論文約640本 (2014年6月現在)

- ざっくりと分類
  - 現象の網羅・類型化
  - 事例研究
  - 言語資源の開発
  - 言い換え認識
  - 言い換え生成
  - 応用技術への適用・導入

言い換え認識

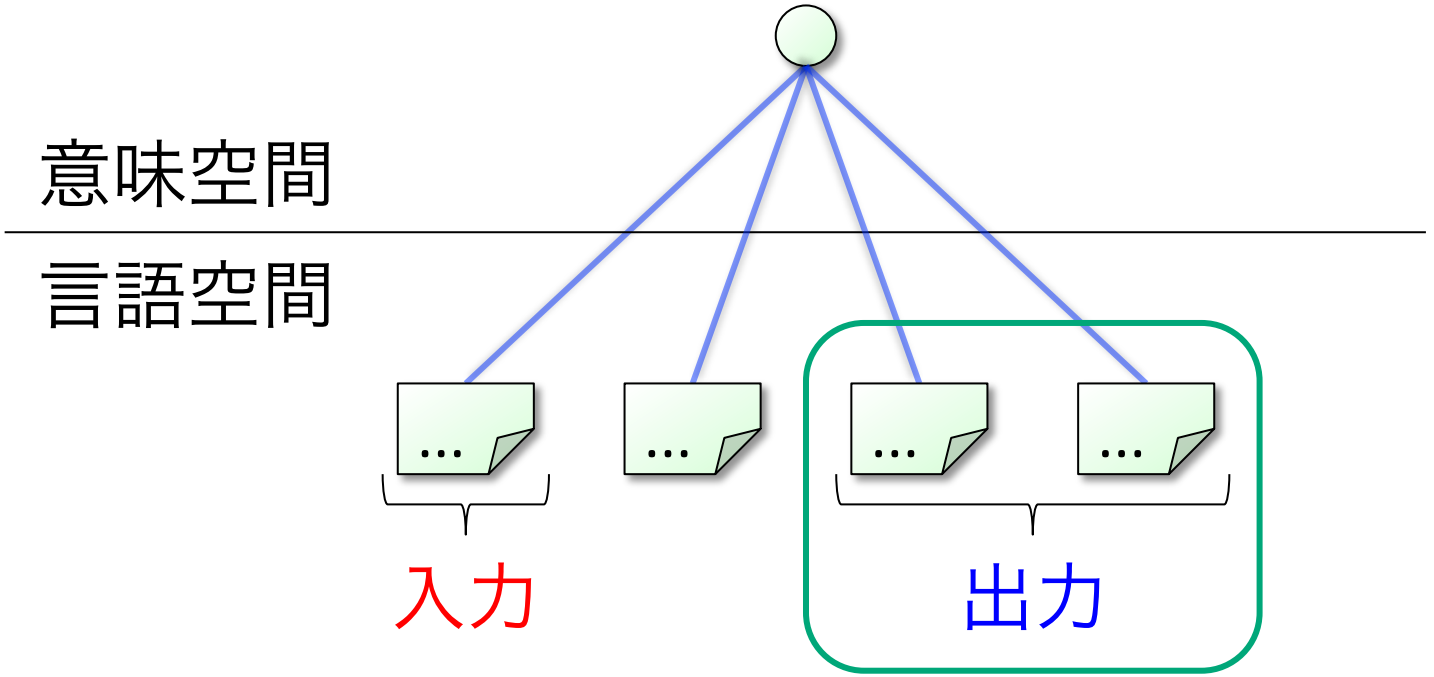
- [入力] 2つの異なる言語表現
- [出力] 同じ意味を表すか否か(あるいはその程度)
- 応用: 情報検索, 質問応答, 複数文書要約, 剽窃検出



出力$_1$ = 同義　出力$_2$ = 同義でない

言い換え生成

- [入力] 言語表現，目的に応じた評価基準
- [出力] 入力と同義で基準を満たす言語表現の集合
- 応用: 平易化，文圧縮，機械翻訳の前処理，折句生成



評価基準を満たす
言語空間の部分空間

言い換え処理に必要な言語資源

- 言い換え処理に必要な知識
  - 語の素性/意味記述
    - e.g., 項構造/意味役割，動詞の使役・受身の可否
  - 表現間の関係に関する知識
    - e.g., 同義の語句の辞書 (言い換え知識)
    - e.g., 派生語，反義語辞書，特質構造 (生成語彙論)
  - 語の共起尤度等の統計情報
  - 全貌はまだまだ不明
- 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
  - 応用: 生成と認識の根幹，あるいは他のタスクにも貢献

どれをやるか?

- 言い換え知識獲得: 有意義な分析例あり
  - Ja: 獲得分の分類・分析 [河合+, 12]
  - En/Fr: ゴールドデータの作成+自動獲得の評価 [Max+, 12]
- 言い換え生成: 難しすぎる
  - 18種類の誤りカテゴリ[藤田+, 03]
    - 8種類約28k個の統語構造変換規則 → 630件の事例
  - 知識の表現方法・規模, 生成手法に強く依存
  - 網羅性の判定が極めて困難
- 言い換え認識: 分析の価値がありそう
  - 関連研究の知見あり
    - En: Microsoft Research Paraphrase Corpus (MSRP) [Dolan+, 04]
    - Ja: NTCIR Recognizing Inference in TExt (RITE2) [Watanabe+, 13]

FY2014の成果

- 言い換え認識に関して分析
- 出発点
  - コーパス: 専用のものはない
  - システム: 専用のものはない
  - 分析メンバ: 経験者はいない
- モノはありませんが道筋は決まりました
  - 客観的かつ精密な評価のためのシナリオ
  - エラー分析に適したコーパスの仕様の整理
    - 一部について実行可能性を調査
    - オープンクエスチョンもあるが…

関連研究のおさらい

関連研究のおさらい: MSRP

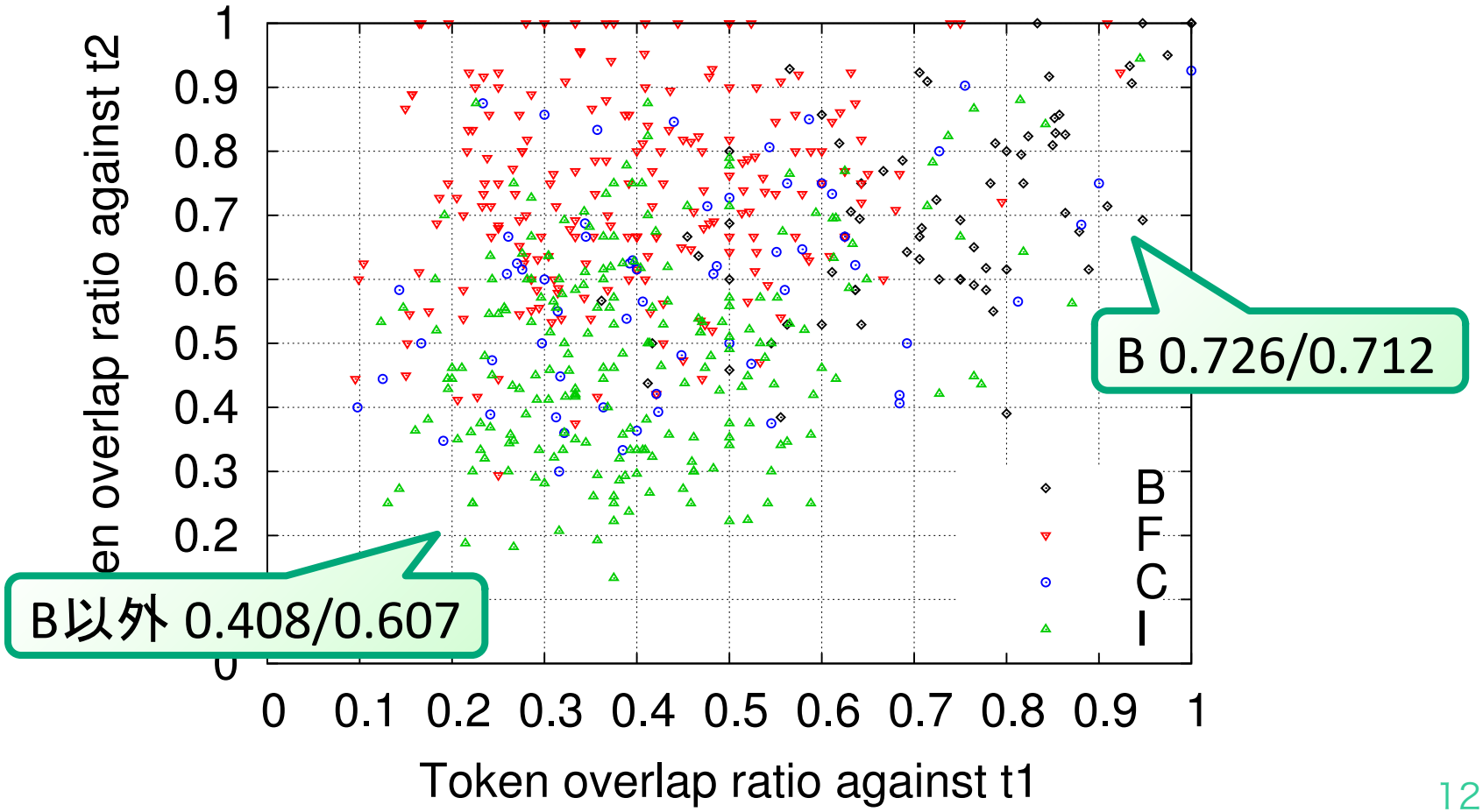
- 英語の言い換え認識評価用データ [Dolan+, 04]
  - コンパラブルな記事中の編集距離8-20の文の対
  - 正例1147件, 負例578件
  - state-of-the-artのF値: 84.1
    - 複数のMT自動評価尺度のスタッキング [Madnani+, 12]
- 評価用データが適切でない
  - 実世界の問題の分布を反映していない [Xu+, 14]
    - cf.「同義」としか言わないベースライン: F値79.9
    - state-of-the-art も実はあまり解けていない
  - 正解ラベルが誤っている場合もある
  - 簡単すぎる? トークン重複率: 正例0.715, 負例0.600

関連研究のおさらい: RITE-2

- 日本語の含意関係認識用データ [Watanabe+, 13]
  - Wikipedia 中のキーワード検索結果から人間が抽出
  - 4分類: B (言い換え), F (一方向), C (矛盾), I (関係なし)
  - state-of-the-artのF値: 69.3
    - 文字の重複に基づく手法 [Hattori+, 13]
- 評価用データが適切でない?
  - 難しすぎる: どこまで解けているのか? [Kaneko+, 13]
  - 逆に簡単すぎる?: e.g.,トークン重複率
- リッチな言語資源がまだ活かせていない
  - 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

RITE-2の評価用データのトークン重複率

- 共通トークン数 / 片方のテキストのトークン数
  - 本来の重複傾向と標本選択バイアスは分離できないが…
  - t_1に対する重複率r_1と正解システム数の相関: 0.771



比較的よく解けている事例

- ID=242, B, r_1=1.00: B(17), F(3), 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, r_1=1.00: B(15), F(4), C(1), I(1)
  - 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, r_1=0.79: B(12), F(2), C(0), 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, r_1=0.75: B(10), F(2), C(4), I(5)
  - 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。

false negative

- ID=292, B, r_1=0.62: B(4), F(10), C(1), I(6)
  - 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, r_1=0.56: B(4), F(3), C(1), I(12)

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, r_1=0.42: B(2), F(7), C(0), 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。

false positive

- ID=20, C, r_1=1.00: F(1), B(12), C(3), I(4)

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。

- ID=65, I, r_1=0.94: F(5), B(11), C(4), I(1)

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。

- ID=91, F, r_1=0.91: F(4), B(14), C(0), I(3)

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。

関連研究のおさらい: RITE-2 (contd.)

- 複数の部分問題を含む複雑な事例 [Kaneko+, 13]
  - エラー分析の非生産性
    - 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
    - 分析者の主観に依存し，情報を共有しづらい
  - 部分点が与えられない
- 事例の偏り
  - トークン重複率の傾向が「浅い」アプローチを奨励
  - 負例は十分に紛らわしいか? (cf. WSC [Levesque, 11])
    - 表層的に，というだけではなく
- 数が少ない: 正例70件

エラー分析に向けてのシナリオ

エラー分析に向けてのシナリオ

- 第1段階　評価に適した言い換えコーパスの構築
  - 3つの要件: 自然な分布，正負例のバランス，プリミティブ
- 第2段階　必要な知識・機能の列挙 [Sammons+, 10]
  - (人間の)判断に必要な知識・処理のインスタンス
    - cf. 人間のプロセス ≠ 理想的なシステム
- 第3段階　既存の技術の客観的評価と課題の提言
  - 上記の評価データに基づいて手法をプロファイリング
  - 語彙資源等の外的な評価 [柴田+, 15(本WS)]
  - ホワイトボックス，グラスボックス評価は開発者に任せる

第3段階 (エラー分析のGoal)

- 既存の手法に依存するのは危険
  - cf.「複数の手法を比較して...」
  - 問題の全体像は(現在の)解き方とは独立
    - 新しい(パラダイムの)手法のエラー分析は一からやり直し!?
    - 言い換え認識の場合はまだstate-of-the-artがない
- 他人のシステムの分析には責任を持てない
  - ホワイトボックス，グラスボックス評価 → できない
  - オラクル調査 → できない
- 手法・システムをプロファイルする
  - どんな部分問題がどれくらい解けている/いないか
  - 解くべき問題をなるべく客観的に分類しておく

第2段階 (エラー分析=解くべき問題の分析)

- エラー分析の方法論そのものを疑う
  - 出てきたエラー(だけ)を分析することの不十分性
    - 表面的には見えないエラー (まぐれ正解)
    - テストデータの十分性
  - 分析の方法論や分類基準の不安定性
    - 同じ人でも見るたびに解釈が異なる
      - 色々な方々「エラー分析しんどい」
    - 異なる人だともっと異なる(に違いない)
      - 誤り分析のガイドラインを作ったとして，追従できるか?
      - 複数人のアノテーション結果の統合 [新納+, 15(本WS)]
      - 藤田早苗さん「自分の感覚にもっとも合ったエラー分類を参考にするのがいいと思います。（みんな結局は自分のエラー分析に基づいて、次にやるべきことを考えると思います）」

誤りの分類体系の例 (1)

- 省略解析 [飯田+, 15(本WS)]

特徴	事例数
アノテーションの誤り・問題	15
機能語相当表現へのアノテーション	10
「いわれた」のような外界照応の問題と混在	9
名詞+” だ” の格要素	9
離れた位置（1 文前 or 2 文前）に先行詞が出現	6
ガ格で先行詞が述語より後に出現	5
名詞句チャンキングの誤り	4
丸括弧の問題	4
ニ格で先行詞が述語より後に出現	3
ひらがな表記が影響	2
ニ格の解析誤り	2
文末の名詞句が先行詞となる	2
その他	29

誤りの分類体系の例 (2)

- WSD [白井+, 15(本WS)]

- 手法の問題
  - 教師あり機械学習に基づく手法の問題
    - 訓練データの不足 (27)[0.134]
      - 他に手がかりなし (21)[0.104]
    - 素性抽出が不適切 (2)[0.010]
      - 意味クラスの抽象度 (5)[0.025]
      - 助詞の取り扱い (10)[0.050]
      - 格の交替の取り扱い (3)[0.015]
      - 連体修飾の取り扱い (8)[0.040]
      - システムのバグ (3)[0.015]
    - 有効な素性の不足 (7)[0.035]
      - トピック素性 (10)[0.050]
      - 長いコロケーション (2)[0.010]
      - 間接的な係り受け (3)[0.015]
      - 既存の素性の組み合わせ (7)[0.035]
      - 文脈に出現する語の語義 (2)[0.010]
      - 語釈文と文脈の関連性 (3)[0.015]
      - 照応・省略解析 (3)[0.015]
    - 素性のコーディングが困難 (1)[0.005]
      - 文の解釈 (20)[0.100]
      - 文脈の解釈 (18)[0.009]
    - 学習アルゴリズムの問題 (14)[0.070]
    - 過学習
  - 辞書の用例に基づく手法の問題
    - 文間類似度の不備
      - 類似度が低すぎる (7)[0.035]
      - 類似度が高すぎる (20)[0.100]
      - 表層的には似ていない (6)[0.030]
    - システムのバグ (1)[0.005]
    - タイブレークが不適切 (1)[0.005]
  - 辞書の文法的制約に基づく手法の問題
    - 文法的制約が緩い (7)[0.035]
    - 規則の不備 (1)[0.005]
  - 分類器の組み合わせ手法の問題 (14)[0.070]
  - 消去法 (14)[0.070]
- 知識の問題
  - シソーラスの不備 (3)[0.015]
- 前処理の問題
  - 形態素解析の誤り
  - 文節の係り受け解析の誤り (1)[0.005]
- データの不備
  - 正解語義の誤り
    - 訓練データ (15)[0.075]
    - テストデータ (32)[0.159]
- 問題設定の不備
  - 文脈不足 (1)[0.005]
  - 対象語が不適切 (16)[0.080]
  - 熟語・連語として扱う方が適切 (5)[0.025]
  - 人間でも判定が困難
- その他 (1)[0.005]

誤りの分類体系の例 (3)

- 言い換え生成 [藤田+, 03]

トランスファの種類	＜格＞	＜否＞	＜機＞	＜サ＞	＜分＞	＜動＞	＜語＞	＜慣＞	合計
評価事例数	138	75	19	39	20	60	221	58	630
不適格性を含む（修正を必要とする）事例数	137	57	9	35	17	53	172	36	516
(a) 《活用形の誤り》	125	41	3	31	7	43	47	6	303
(b) 《不適格な機能語連接》	42	14	2	3	5		8	4	78
(c) 《格助詞の欠損》		*			6	2			8
(d) 《同じ格要素の重複》				7				4	11
(e) 《節内の格要素と動詞の不整合》	66	*	*	8		28	57	3	162
(f) 《修飾語の重複，競合》				*					0
(g) 《(e) 以外の共起の不整合》				*		3	28	5	36
(h) 《内容語の意味の変化で文の意味が変わる》							30	1	31
(i) 《モダリティの持つ意味の変化で文の意味が変わる》	1	5		3			13		22
(j) 《時間情報が等しくない》	2	1			3				6
(k) 《文体が等しくない》	*		1	*					1
(l) 《すわりが悪い語順》	23	*	*	*	2	7		2	34
(m) 《主題・陳述構造の不整合》	10	1			10	1			22
(n) 《節間，文間の修辞的関係の不整合》	2	4	2						8
その他	38	16	2	7	8	3	19	22	115
(A) 慣用表現・固有表現の誤認	9			1			26	4	40
(B) 辞書特有のメタ表現によるノイズ							18	20	38
(C) 形態素・構文解析の誤り	7	5	5	1			22	1	41
(D) 言い換えエンジンの書き換え操作の誤り	8	1	1		1	1	1	2	15

第2段階 (エラー分析=解くべき問題の分析)

- 解くべき問題をあらかじめ確認しておく
  - 正しい出力を得るために必要な知識・処理を列挙する
    - インスタンスの列挙，類型化・体系化 [Sammons+, 10]
    - 既存のオントロジ/タイポロジから出発
    - OntoNotes方式 [Hovy+, 06] で収斂
  - 他のタスクでも可能かも?
    - 個々の誤答の理由を，あらかじめ説明しておく
      - 賀沢さん「“～もらおう” が “考える” の語義をやや示唆」

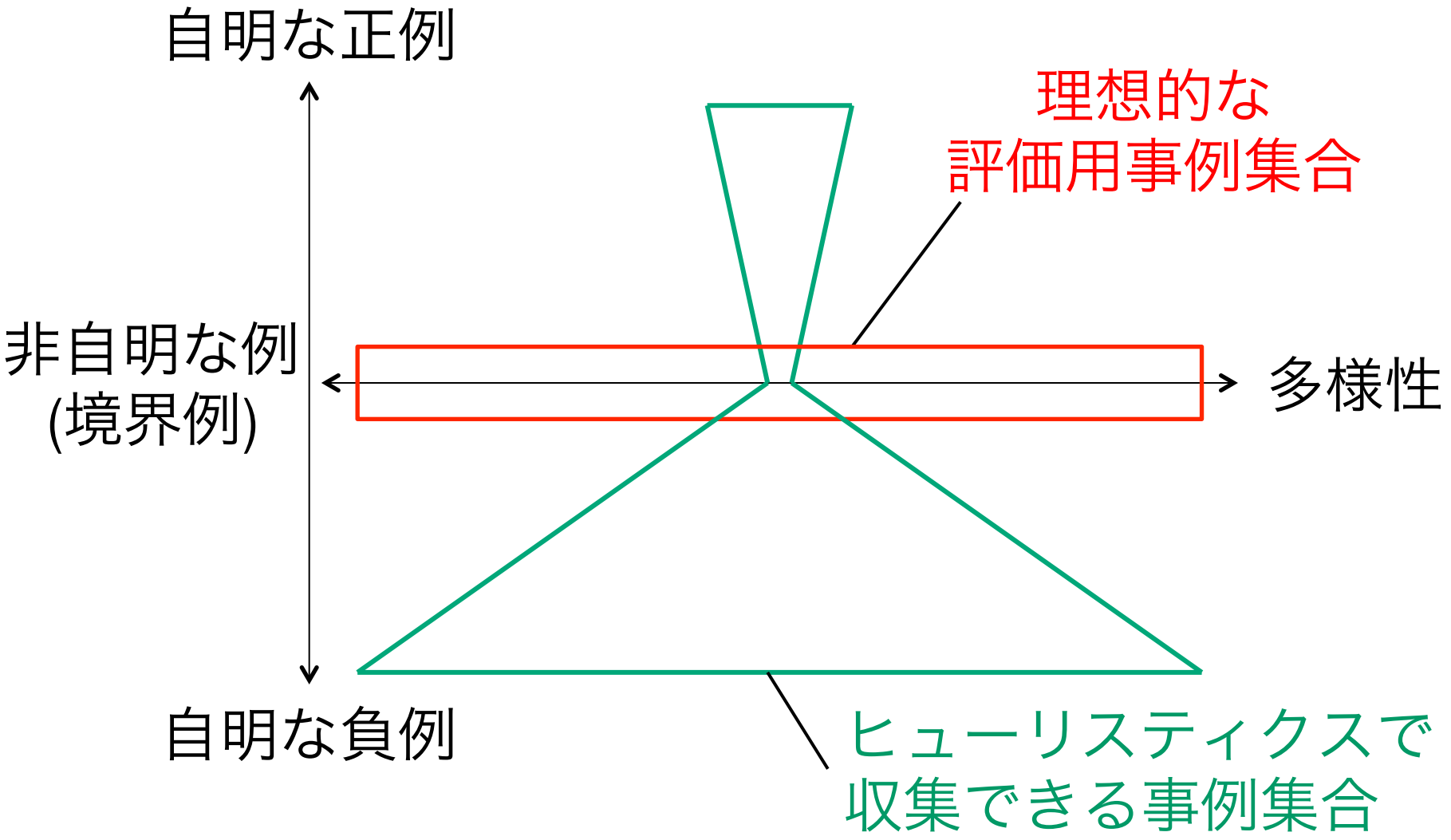
第1段階 (エラー分析を可能にするデータ)

- 評価に資するコーパスが満たすべき3つの要件
  - 分布の自然さ: お手上げ状態
    - 本当に解きたい問題の分布を反映したサブセット
      - サンプルバイアスの例「編集距離8-20」 [Dolan+, 04]
    - 応用ごとに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
      - e.g., トークン重複率で解ける
    - 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]
    - Ref. WSC [Levesque, 11]
  - プリミティブさ: 担保可能
    - 独立な部分問題への分割と分類
    - 公平な評価: 惜しい誤答に部分点，まぐれ正解を減点

評価用データの作り方

- 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)
- どうすればいいんだ!?
  - 言い換え認識
    - 任意のテキスト対はほぼ間違いなく言い換えではない

収集したい言い換え事例の範囲



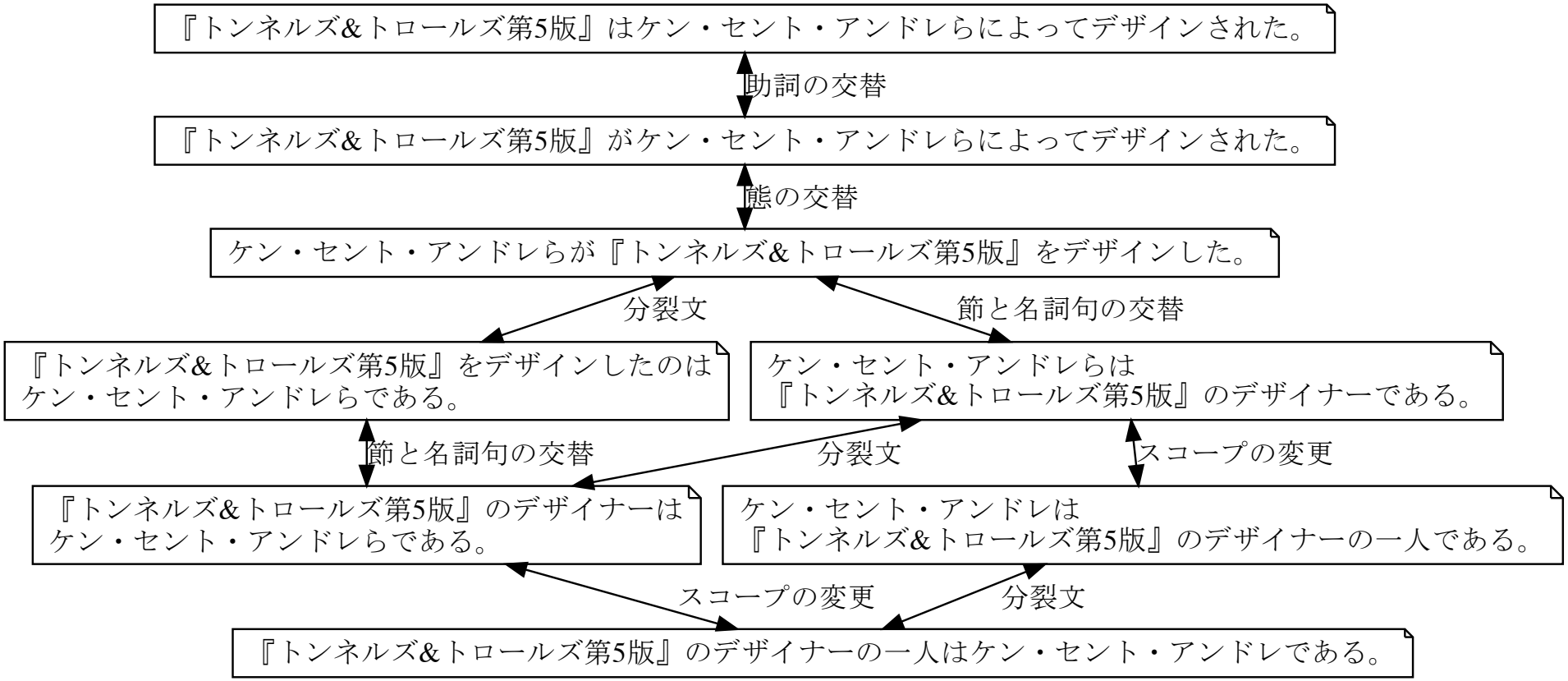
事例の分解可能性

事例の分解は本当にできるのか?

- RITE2のユニットテストデータ [Kaneko+, 13]
  - 本活動と同じ意図で作成されたデータ
  - 含意か否かの分類 (言い換えではない)
  - まだ複雑
- さらに分解
  - 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

- 163事例 → 306件のプリミティブな事例
  - 内訳
    - 分解なし: 108事例
      - 言い換え: 58件
      - 非言い換え: 45件 (e.g., 文の一部の抽出)
    - 分解あり: 60事例 → 203事例
      - 言い換え: 156件
      - 非言い換え: 47件
  - 42種類+その他

言い換えの種類の分布

言い換えの種類	分解済	新規獲得	合計
名詞/名詞	1	7	8
名詞句/名詞句	0	2	2
動詞/動詞	0	6	6
動詞/動詞句	1	2	3
動詞句/動詞句	1	2	3
副詞/副詞	1	1	2
略記	0	1	1
表記の揺れ	0	2	2
助詞の交替	2	31	33
助動詞	0	4	4
テンス・アスペク	0	2	2
機能表現	0	3	3
複合名詞化	1	1	2
同格表現の異形	1	0	1
括弧・同格	1	0	1
括弧の付加/削除	0	5	5
並列名詞句の入れ替え	2	1	3
並列動詞句の入れ	0	2	2
格要素の語順の変	4	9	13
数量詞の移動	0	1	1
主題の交替	9	10	19
態の交替	5	6	11
相互格の交替	2	0	2
機能動詞構文	1	4	5
動詞句/名詞句	0	1	1
文法カテゴリを変える言い換え	0	3	3
所有–存在	0	1	1
地名–存在	1	1	2
分裂文	0	2	2
節/名詞句	0	3	3
節の統合/分割	1	0	1
節の連体修飾節化	2	1	3
節をまたぐ言い換え	0	2	2
文の統合/分割	0	4	4
共参照表現による	3	5	8
コピュラ文の主語	1	3	4
自明要素の明示/暗示	15	9	24
説明の省略	2	3	5
数量詞の省略	0	1	1
非制限的説明の除去	0	2	2
スコープの変更	0	1	1
句読点	0	4	4
未分類	1	8	9
合計	58	156	214

内容語句

スタイル

機能表現

複合名詞

並列要素

節内構文

節間構文

呼応

その他

自明要素の明示/暗示

- ID=bc580-0-6:

『ステンカ・ラージン』はウラジミール・ロマシコフが監督、ワシーリ・ゴンチャロフが脚本の映画だ。

『ステンカ・ラージン』はウラジミール・ロマシコフが監督、ワシーリ・ゴンチャロフが脚本で制作された映画だ。

- ID=bc-160-2-1

カルマ・カギュ派が、化身ラマ制度を初めて法主の選任に採用した。

カルマ・カギュ派が、化身ラマ制度を初めて採用した。

予稿に書いてないこと

藤田の個人的試みと私見

ボトムアップな言い換え事例収集

- ある範囲のテキストをとことん言い換える (内省)
  - 仮説: 100人集めればある程度の網羅性を担保できる
  - パイロット作業
    - BCCWJから言い換え元の文をサンプル
    - ひたすら言い換え → 150事例/5時間 (ペースはほぼ一定)
      - trivial なものも結構含む
      - minimal pair となる負例は別途要作成
  - 宮尾さん「人間の限界はたかがしれている」

応用指向で問題を定義する

- 分布の自然さはおいといて応用への貢献を目指す
  - 各タスクでどんな種類の言い換えを解きたいか?
    - WSD in MT [藤田さ+, 15(本WS)]
    - 言い換えとそれ以外の現象の線引
      - 多くの方々「言い換え大事だよね〜」本当に言い換え?
      - cf. 一昔前の「慣用句」
    - 応用タスクの例: 含意関係認識, 複数文書要約, etc.
  - それが解けた時にどれくらいインパクトがあるか?
    - どこにでも存在すると言うけれども ...
    - e.g., RITE-2 w/ 言い換えフレーズ対250万対 → 全然当たらん
    - e.g., SMTのOOV解消 → X%

FY2014の活動のまとめ

- 出発点
  - コーパスなし，システムなし，経験者なし
- 成果物
  - 客観的かつ精密な評価のためのシナリオ
  - エラー分析に適したコーパスの仕様の整理
    - 一部について実行可能性を調査
    - オープンクエスチョン: 自然な分布をどう近似するか
      - 全体像は無視して特定の部分問題だけ優先的に潰す
        - ✓ 応用側で解きたい部分問題
        - ✓ 頻出する部分問題