語彙的対応関係の一般化に基づく言い換え知識の拡張

藤田 篤
Pierre Isabelle


情報通信研究機構


National Research Council Canada

今日の話題

- 与えられた言い換え知識の自動拡張
  - 言い換え知識: 言い換え関係にある表現の対の集合
  - 提案手法
    - 関係ありそうな語の対の抽象化

amendment of regulation ⟷ amending regulation

X:ment of *Y*:ϕ ⟷ *X*:ing *Y*:ϕ

    - 大規模単言語コーパスからの具体例の抽出

investment of resources ⟷ investing resources

recruitment of engineers ⟷ recruiting engineers

  - 実験結果
    - 規模: コーパスの規模によるが100倍のオーダーは達成可能
    - 精度: ナイーブな使い方でも従来手法と同程度以上

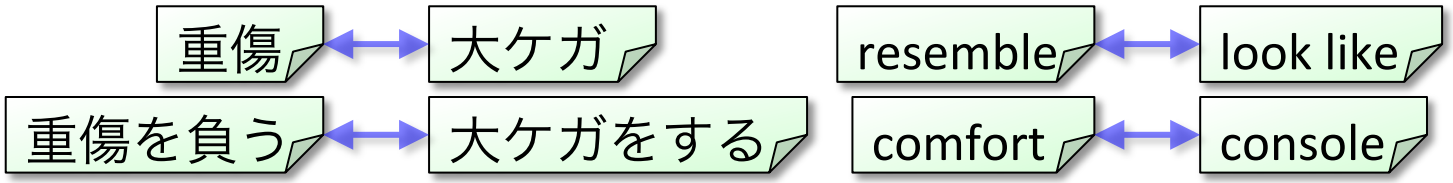
背景と動機

言い換え:概ね同じ意味内容を表す言語表現

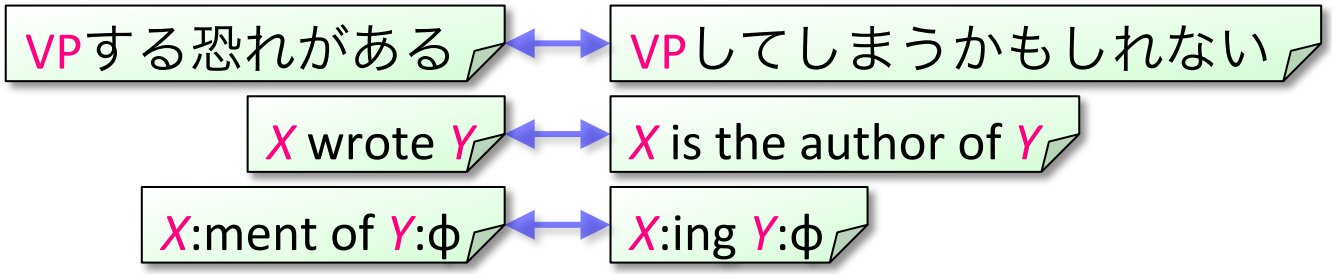
- 語や語の列
  - The roof looks like a prehistoric lizard's spine.
  - The roof resembles a prehistoric lizard's spine.
- 節内の構造変換
  - Employment showed a sharp decrease.
  - Employment decreased sharply.
  - The car collided with the bicycle.
  - The car and the bicycle collided.
- 複数の節にまたがる構造変換
  - It was his best suit that John wore to the dance last night.
  - John wore his best suit to the dance last night.

言い換え(生成)に必要な知識

- とくに言い換え知識
  - 言い換え関係にある表現の集合(一般に対)の集合
    - 何らかの文脈で同義かつ置き換え可能
  - 一般に文よりも小さなテキスト断片
    - 表層表現 (構造情報を含む場合も)



    - パターン，テンプレート



言い換え知識の自動獲得の研究動向

- 獲得元
  - (シソーラス)
  - 単言語(シングル)コーパス
  - 単言語パラレルコーパス
  - 単言語コンパラブルコーパス
  - 異言語パラレルコーパス (対訳コーパス)
- 手がかり
  - 周辺文脈 [Harris, 54][Lin+, 01]
  - 共通の翻訳 [Barzilay+, 01][Bannard+, 05][Callison-Burch, 08]
  - 概念，トピック，用語 [Shinyama+, 02][Hashimoto+, 11]

単言語シングル + 分布仮説

- 分布仮説 [Harris, 54]
  - “類似する文脈でよく使われる表現は似た意味を持つ”
    - 文脈素性: 左右の隣接 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

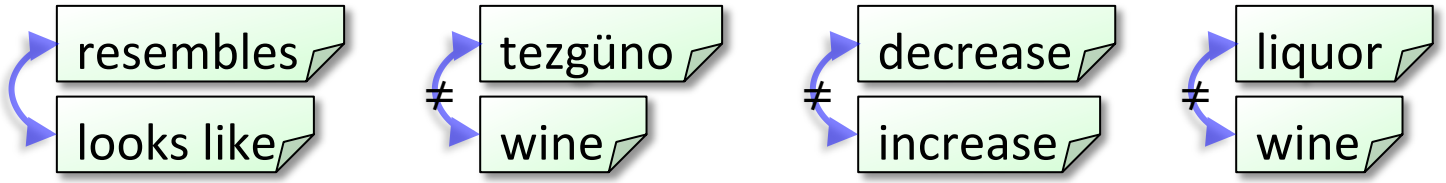
A bottle of tezgüno is on the table

Everyone likes tezgüno

Tezgüno makes you drunk

We make tezgüno out of corn

  - 同義とは限らない: e.g., 反義語, 上位下位語

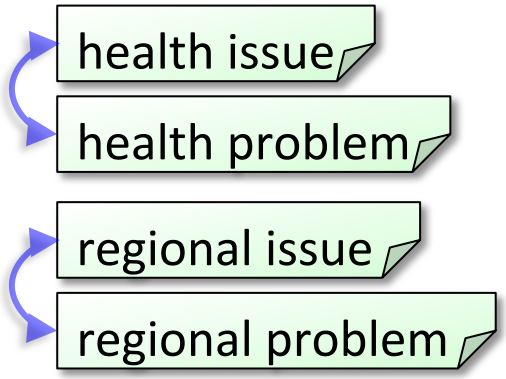


異言語パラレル + アラインメント

- 翻訳を意味表現として用いる [Bannard+, 05]
  - 分布仮説における「文脈」よりも信頼できる
    - 多義性の問題は残る
  - 対訳コーパスから自動的に獲得可能
    - 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

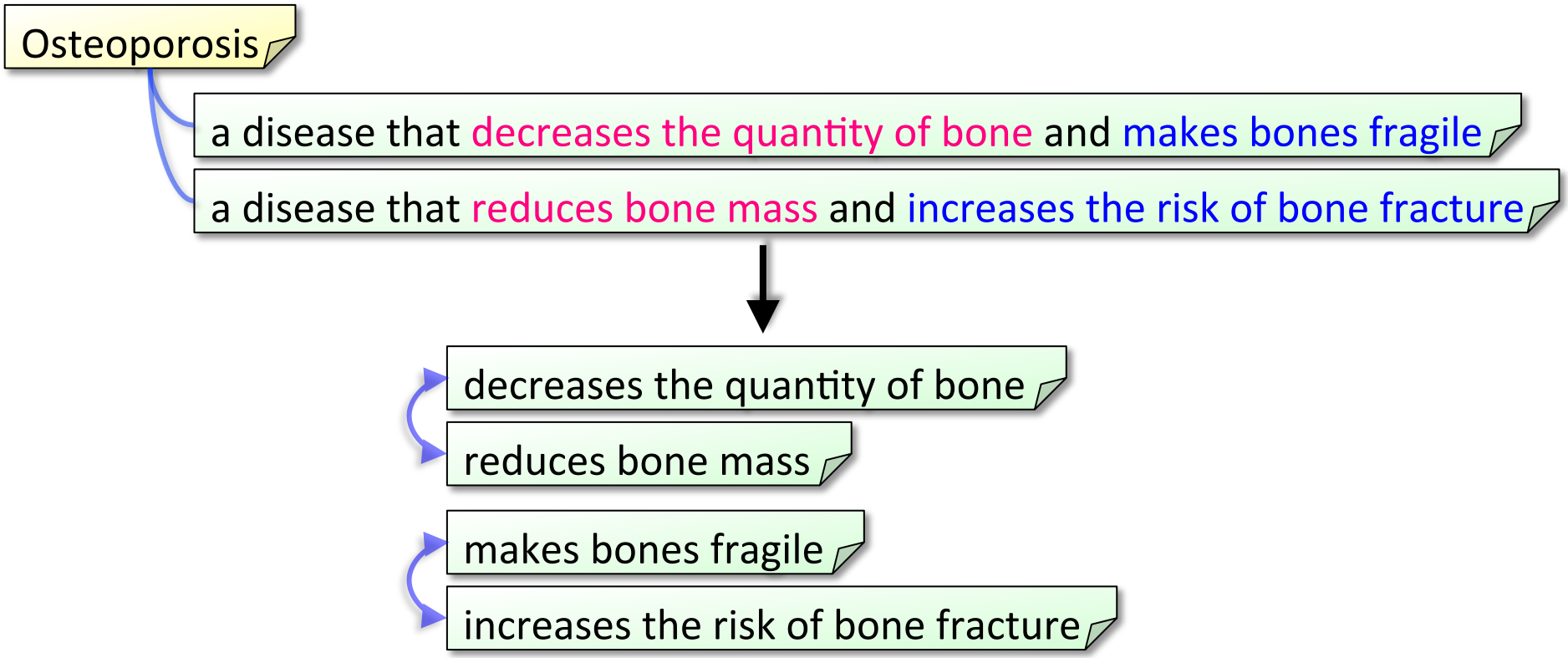
自動的に学習したフレーズテーブル

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



単言語コンパラブル + アラインメント

- 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]
  - コンパラブルコーパス << シングルコーパス



カバレージと精度の両立が課題

- 単言語コーパス + 分布仮説
  - Pro. 大規模コーパス → 高カバレージ
  - Con. 周辺文脈の類似性は弱い手がかり → 低精度
    - e.g., 反義語，上位下位語も似た文脈で使われる

- パラレル/コンパラブルコーパス + アラインメント
  - Pro. 文単位の同義性/類義性 → 高精度
  - Con. 規模が限られる → 低カバレージ

品質を保ったまま知識を拡張する

与えられた言い換え知識の拡張

- タスクの定義
  - 入力: 高品質な言い換え知識
  - 出力: 新たな言い換え知識
- 提案手法
  - 関係ありそうな語の対の抽象化

amendment of regulation ⟷ amending regulation

X:ment of *Y*:ϕ ⟷ *X*:ing *Y*:ϕ

  - 大規模単言語コーパスからの具体例の抽出

investment of resources ⟷ investing resources

recruitment of engineers ⟷ recruiting engineers

手続き

S_{Seed}：種となる言い換え表現対の集合

airports in Europe ⇔ European airports
amendment of regulation ⇔ amending regulation
should be noted that ⇔ is worth noting that

Step 1. 言い換えパターンの獲得

言い換えパターン

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

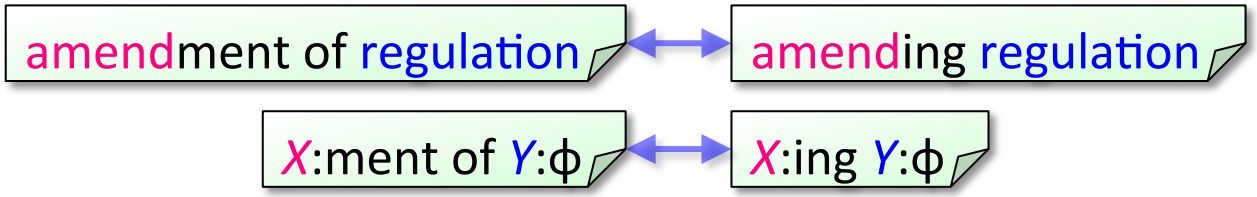
Step 2. 新規言い換え表現対の収集

S_{LV}：新規言い換え表現対の集合

cohesion in Europe ⇔ European cohesion
democracy in Europe ⇔ European democracy
increase in Haiti ⇔ Haitian increase
transportation in suburb ⇔ suburban transportation
economy in Uruguay ⇔ Uruguayan 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

Step 1. 言い換えパターンの獲得

- 種となる(高品質な)言い換え表現対の集合 (S_{Seed})
  → 言い換えパターンの集合
  - 両辺において対応関係にある語を変数化
  - 作業仮説: 言い換え表現対の各辺にあり,
    同じ語幹を持つ語の対は, 特定の(意味的な)関係を持つ



対応関係のある語の同定

- 語幹を厳密には同定せず，文字ベースの候補生成
  - (“that”, “this”) → (“*X*:at”, “*X*:is”) / “th”
  - (“oblige”, “force”) → (“oblig:*X*”, “forc:*X*”) / “e”
  - (“amendment”, “amending”) → (“*X*:ment”, “*X*:ing”) / “amend”
  - (“unclear”, “clear”) → (“un:*X*”, “ϕ:*X*”) / “clear”

- 接辞パターン候補のフィルタリング
  - 長さ *k* 以上の語幹 *n* 種類に対して観察された接辞パターン候補のみを残す [Gaussier, 99]
    - *k*: 共通部分の最小文字数 (5)
    - *n*: 語幹の種類 (2)

自動的に同定した語対の例

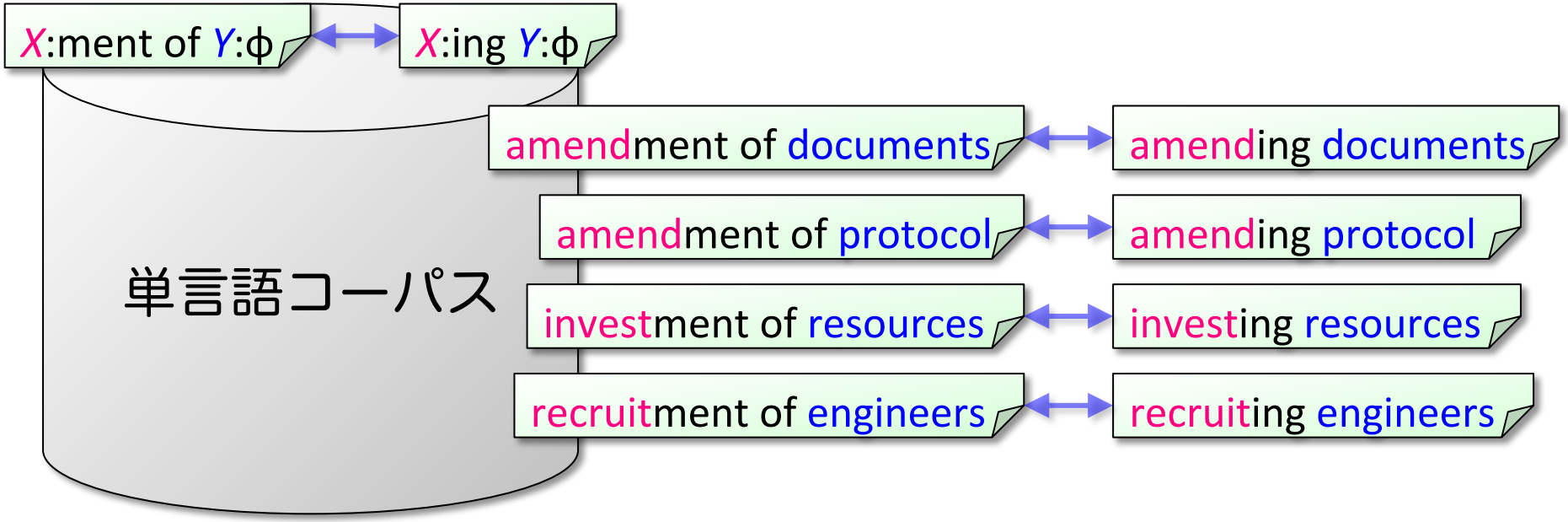
- *X*:ϕ in *Y*:ϕ ↔ *Y*:an *X*:ϕ
  - (“Haiti”, “Haitian”), (“Tibet”, “Tibetan”), (“Uruguay”, “Uruguayan”), (“Chicago”, “Chicagoan”), (“Elizabeth”, “Elizabethan”), (“suburb”, “suburban”)

- *X*:ment of *Y*:ϕ ↔ *X*:ing *Y*:ϕ
  - (“development”, “developing”), (“treatment”, “treating”), (“payment”, “paying”), (“employment”, “employing”), (“investment”, “investing”)

- *X*:able *Y*:ϕ ↔ *Y*:ϕ which can be *X*:ed
  - (“adjustable”, “adjusted”), (“reportable”, “reported”), (“playable”, “played”), (“diversifiable”, “diversified”), (“recordable”, “recorded”)

Step 2. 新規言い換え表現対の収集

- 言い換えパターンの集合 + 単言語コーパス
  → 新規言い換え表現対の集合 (S_{LV})
  - 各パターンの両辺にマッチする表現を抽出
    - 両表現の存在性を重要視
  - 各対の信頼性を文脈類似度で定量化



信頼性: 表現の存在性と表現対の類似性

- 表現の存在性: コーパス中に一定回数出現
- 表現対の類似性: 文脈類似度
  - 文脈素性: 隣接する左右トークンn-gram [Marton+, 09]
    - cf. bag-of-words: 安価だがノイジー
    - cf.依存構造上の隣接要素: 正確だが高価

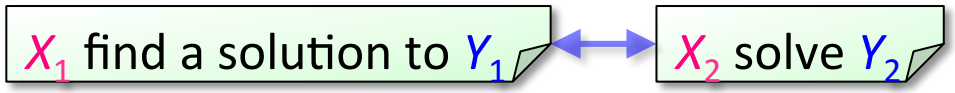
There have been many approaches to compute the similarity between words based on their distribution in a corpus.

to ___ between

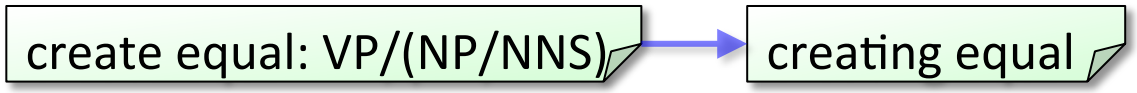
  - ベクトルの比較: コサイン類似度, Jaccard係数, etc.

先行研究 (1)

- 表現の外側に変数を制約として付加
  - 元の表層表現のスコアリングのために変数を付加 [Lin+, 01][Ravichandran+, 02][Szpektor+, 04]
    - Similarity(“find a solution to”, “solve”) =def GeoMean(Similarity(X_1, X_2), Similarity(Y_1, Y_2))



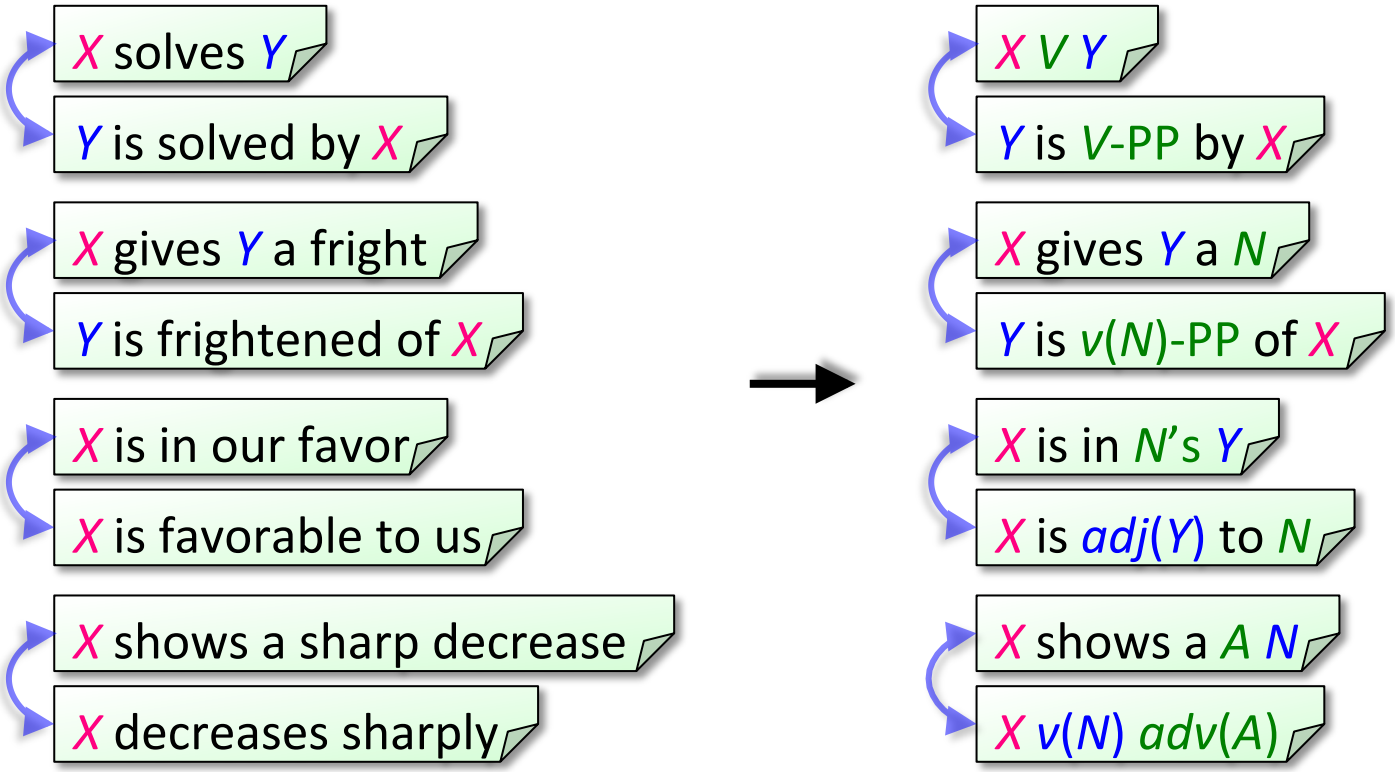
  - 両者が置換可能な条件を特定 [Callison-Burch, 08][Zhao+, 09]
    - “create equal” が右側のNPを含まないVPで, そのNPがその右側の複数形の名詞句(NNS)を含まないならば, “creating equal” に置換可能



  - 集合拡張の諸手法も対象の外側の抽象化 (≒ 文脈類似度)

先行研究 (2)

- 人間による帰納 [Jacquemin, 99][Fujita+, 07]
  - 言い換えにおける派生語間の関係に着目
  - 高コスト，言語依存
    - 関数の実体となる辞書も人間が編纂



提案手法の特徴

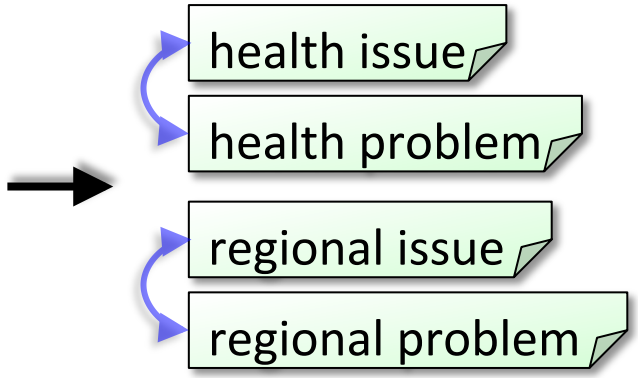
- 全自動・教師なし
  - ごく少数のパラメタ (カバレージと精度のバランサ)
- 種々の語彙的対応関係をカバー
  - 派生語
  - 語形変化 (e.g., 数, 性, 格)
  - 異表記 (e.g., -nize/-nise)
- 単言語コーパスが大きいほどレバレッジが効く
  - 対訳コーパスを大きくするよりも安価

英語の言い換え知識獲得実験
規模・精度

Step 0. 種となる言い換え知識の獲得

- 異言語パラレルコーパス(対訳コーパス)
  → 種となる(高品質な)言い換え表現対の集合 (S_{seed})
  - Bilingual pivoting [Bannard+, 05]

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



  - 各種クリーニング [Fujita+, 12]
    - ストップリスト
    - 言い換え確率

$$p(e_2|e_1) = \sum_{f \in tr(e_1) \cap tr(e_2)} p(e_2|f)p(f|e_1)$$

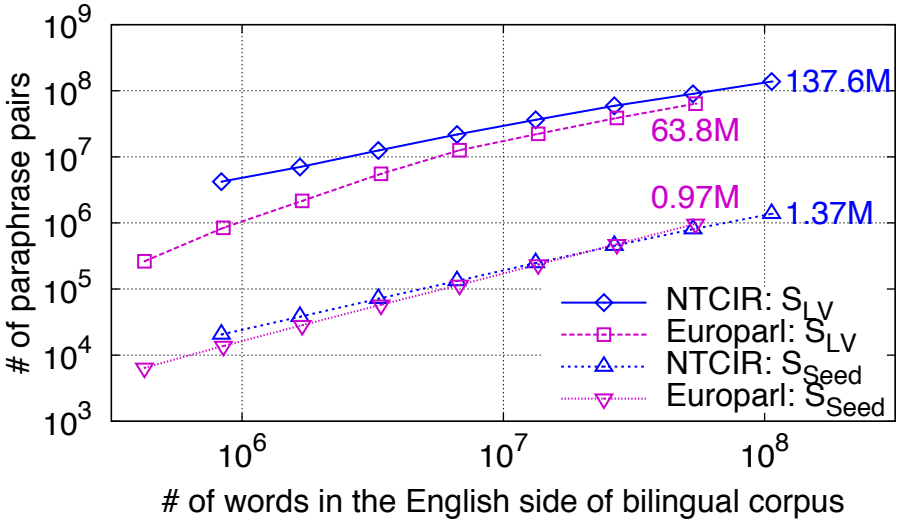
データ

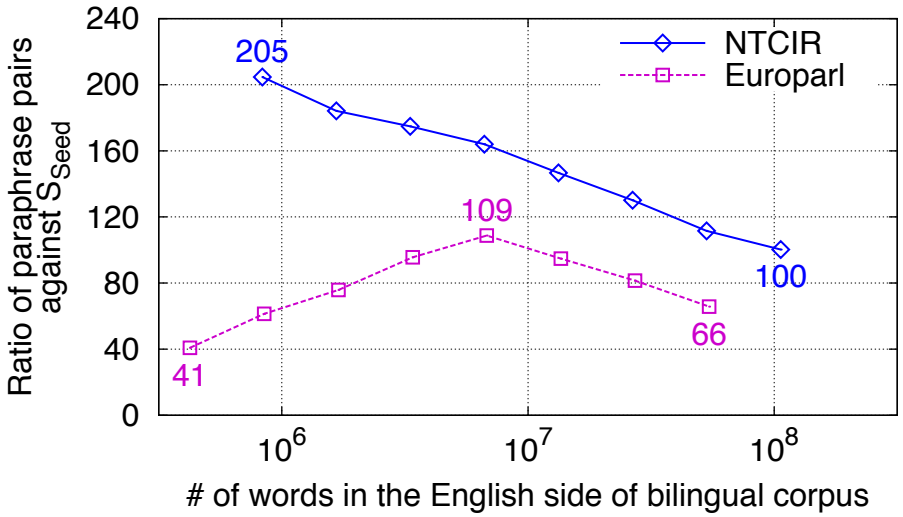
- コーパス (2種類の設定)
  - Europarl (En-Fr)
    - 対訳: Europarl (v7) 2.0M文/55.7Mトークン (en)
    - 単言語: WMT News Crawl (2011-2013) 52.0M文/1.20Bトークン
  - NTCIR (En-Ja)
    - 対訳: NTCIR 特許翻訳コーパス 3.2M文/107Mトークン (en)
    - 単言語: NTCIR 2006-2007年分 39.9M文/1.36Bトークン

使用したツール

- トーカナイザ
  - En, Fr: Mosesdecoder (RELEASE-2.1.1) の tokenizer.perl
  - Ja: MeCab (0.996)の分かち書き
- 単語アラインメント: SyMGIZA++
  - IBM-2 単語アラインメント
- フレーズテーブルの学習: Mosesdecoder
  - Diag-grow-final
  - フレーズ長の上限: 8
  - Significance pruning [Johnson+, 07]

新たに獲得できた言い換え表現対の数

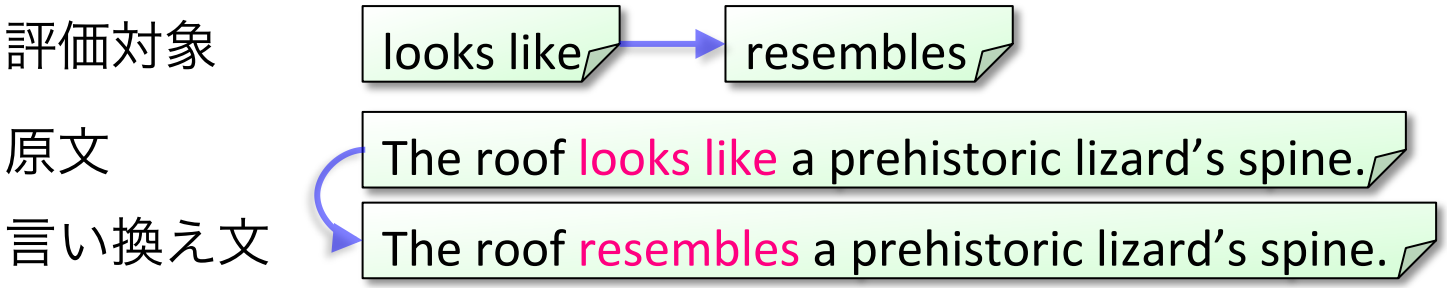




- 顕著に大規模な言い換え知識を構築可能
  - 対訳コーパスが大きいほど
    - 獲得できる言い換え表現対が多い (P_{Seed}, P_{LV} ともに)
    - コーパスのサイズ比が小さいためレバレッジは小さい
      - Europarlの1/16以下を用いた設定では異なる振る舞い
  - 単言語コーパスはもっとある → さらなる大規模化も可能

掘り当てたのは本当に金か?

- 人間による精度の評価: 文脈における置換テスト
  - intrinsicにしたいが
  - 言い換え生成 (ナイーブな置換+LM) [Callison-Burch, 08]



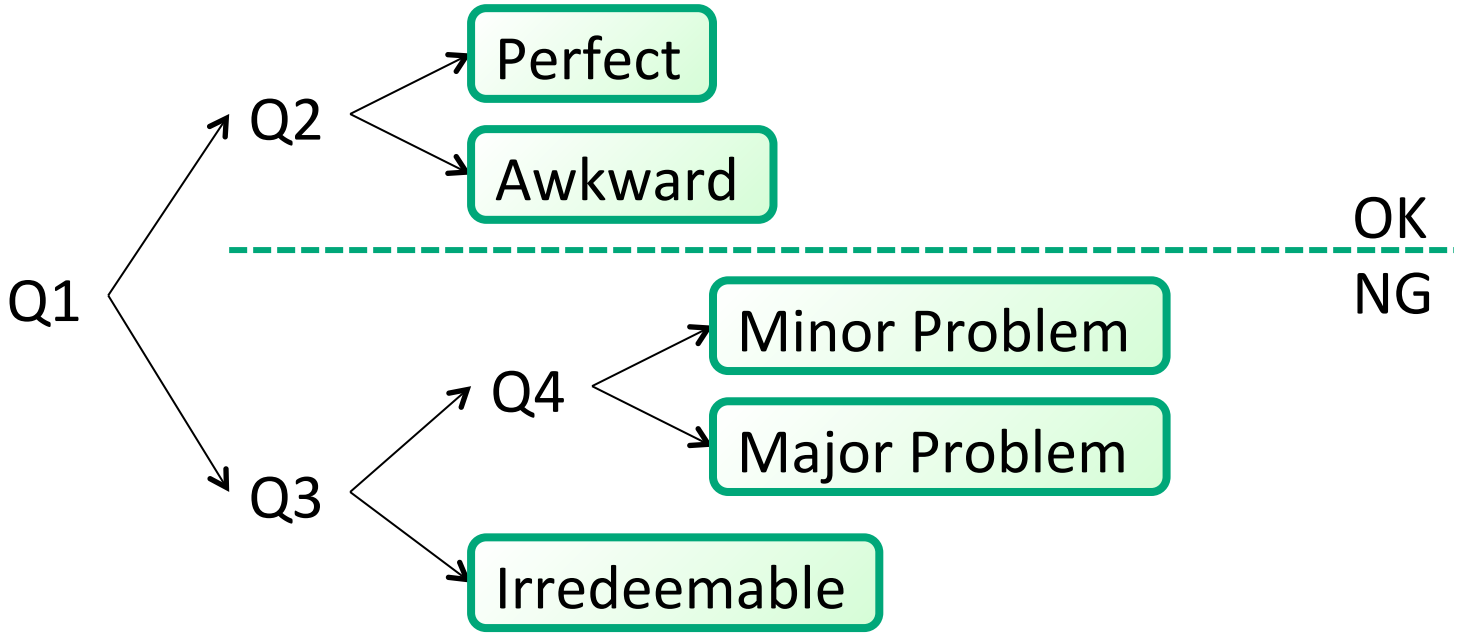
  - プロトコル [Fujita, 13]
    - 同じ箇所に対する複数の言い換え候補を同時に見ながら
    - 2つの基準で
      - まず文法性の評価 (5段階)
      - 次に意味の等価性の評価 (6段階)
    - 各々決定木に従って

評価対象データ

- 候補生成
  - WMT 2011-2013 “newstest” data (10,050文)
    - 制約: 10-30 words (5,850文)
  - Europarl 設定の P_{Seed} (0.97M対) と P_{LV} (63.8M対)
  - 89k箇所に対する1.0M件の言い換え事例
- スコアリング
  - 5-gram LMを用いて各箇所に対するベスト3を選択
- サンプリング
  - 重複のない語句200件 (言い換え文対600件)
    - P_{Seed} 由来: 66件
    - P_{LV} 由来: 534件

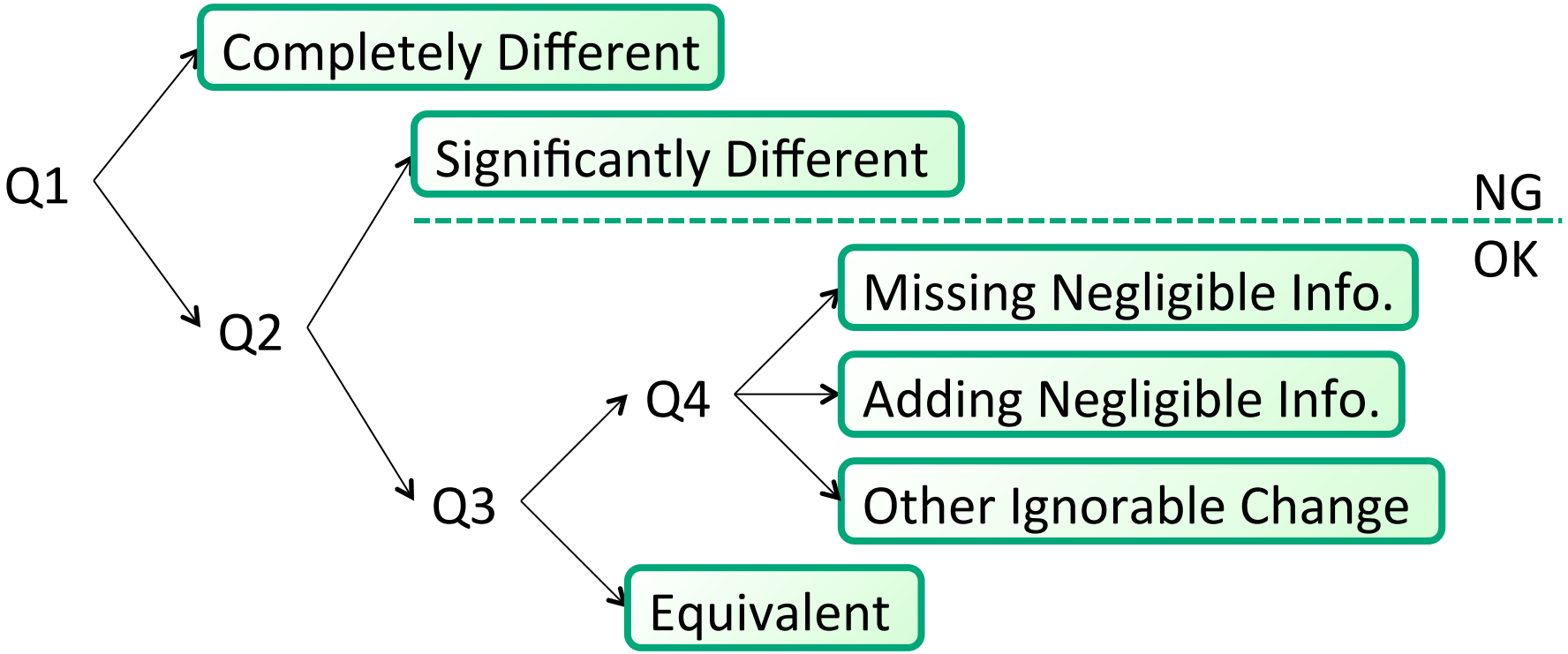
文法性の評価

- 評価対象: 言い換え後の文のみ
- 評価者3名のKappa値
  - 5段階: 0.51 - 0.56 (moderate)
  - 2段階: 0.64 - 0.79 (substantial)



意味の等価性の評価

- 評価対象: 言い換え前の文と言い換え後の文の対
- 評価者3名のKappa値
  - 6段階: 0.27 - 0.35 (fair)
  - 2段階: 0.48 - 0.53 (moderate)



評価結果

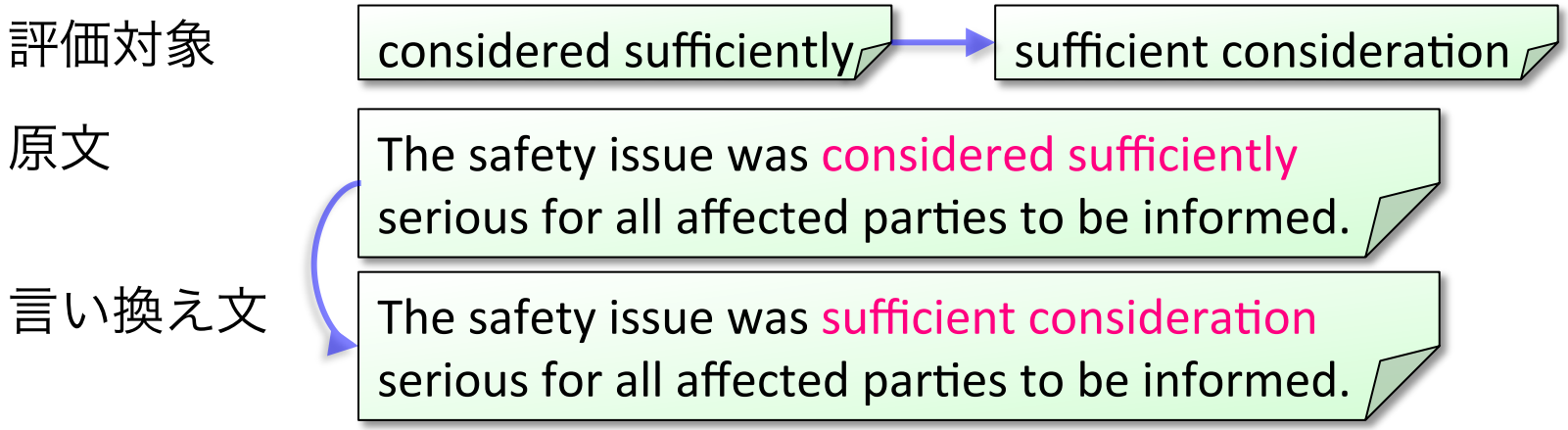
- 精度: 3人の評価者，2段階評価の多数派に基づく

	文数	文法性	意味の等価性	両方
S_{Seed}	66	0.85	0.91	0.76
S_{LV}	534	0.74	0.78	0.59
合計	600	0.75	0.79	0.61

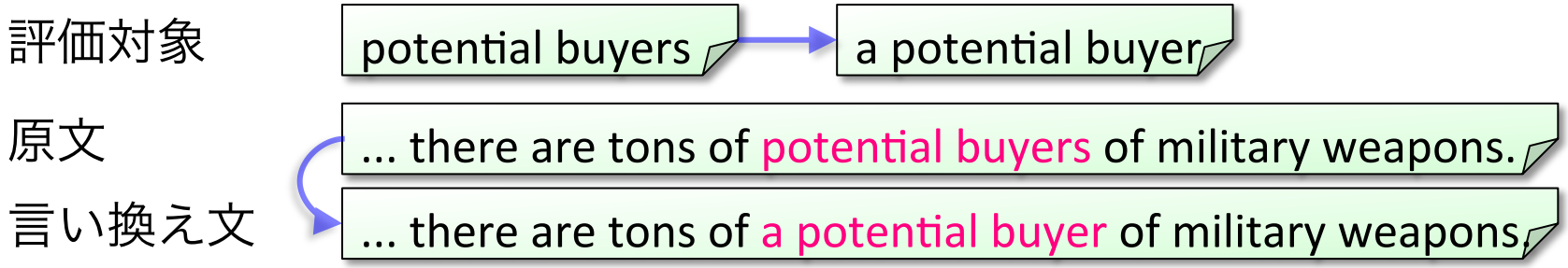
- S_{Seed} 由来の言い換えの精度は十分に高い
  - クリーニングの効果 [Fujita+, 12; 13]
- S_{LV} 由来の言い換えも従来手法と同等以上
  - state-of-the-artにおける値 [Callison-Burch, 08]
    - 評価者，データが異なるので参考まで
    - Europarl (10言語-En) + CCG + LM
    - 文法性 0.68, 意味の等価性 0.62, 両方 0.55

主なエラーの原因

- フレーズ全体のカテゴリ交替



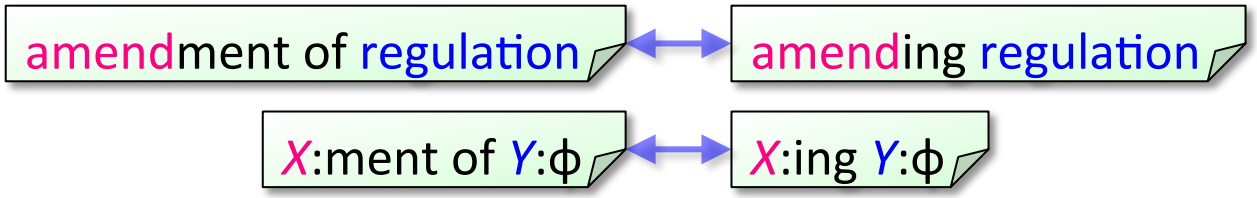
- 数や冠詞などの Agreement 誤り



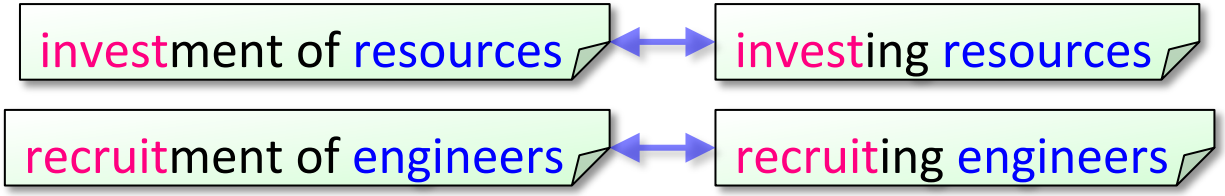
- 言い換え認識の応用タスクでは有用

まとめ

- 与えられた言い換え知識の自動拡張
  - 言い換え知識: 言い換え関係にある表現の対の集合
  - 提案手法
    - 関係ありそうな語の対の抽象化



    - 大規模単言語コーパスからの具体例の抽出



  - 実験結果
    - 規模: コーパスの規模によるが100倍のオーダーは達成可能
    - 精度: ナイーブな使い方でも従来手法と同程度以上