WEB文書にも対応できる日本語異表記の認定基準

黒田 航* 風間 淳一* 村田 真樹*,** 鳥澤 健太郎*

情報通信研究機構 Mastar Project 言語基盤グループ

言語処理学会第16回年次大会口頭発表

2010/03/11, 東京大学本郷キャンパス

目的

- 日本語の異表記 (allography) の認定基準の “標準化” を目指す
  - 異表記の認定基準は言語資源によってばらばら
    - しかも時々直観に反する
  - なおかつ，Webデータの複雑性にも対応可能な基準が必要
- その一環として，大規模な異表記データを開発し ALAGIN フォーラム (http://www.alagin.jp) を通じて配布する
  - 小島ら「機械学習と種々の素性を用いた編集距離の小さい日本語異表記対の抽出」(発表A4-1) のSVM分類器で候補生成，分類の結果を更に人手評価したデータ
  - 規模: 正例約3万，負例約7万

発表の流れ

- 異表記とは何(であるべき)か?
  - 解決すべき問題の定義
    - (1) 異表記対と誤表記対の区別, (2) 異表記と編集距離の近い同義語対の区別が難しい場合がある
  - 解決のための提案
    - (1) 同義条件 (2) 異語条件 (3) 標記の正式性条件の三つを組み合わせた異表記の定義を提唱
    - 特徴: 用途に応じて異表記の認識範囲を変更できる
- まとめ

異表記は何(であるべき)か? 1/3

- 異表記は日本語言語処理では，かなり深刻な問題
  - (i) ひらがな, (ii) カタカナ, (iia) 全角ローマ字, (iib) 半角ローマ字, (iii) 漢字, (vi) 送り仮名の変異の組合わせ
    - 組合わせ爆発が生じる
  - 新規な標記が発明される
    - “ネ申” (“神” の新標記)
- 日本語ほど異表記率の高い言語は稀
  - 異表記の多さはデータスパースネスを悪化させる要因
  - 自動獲得されたトークンの標記の標準化が必要

異表記は何(であるべき)か? 2/3

- 狭義の異表記
  - (仙台, 仙臺), (渡辺, 渡邉)
  - (一円, １円)
  - (Python, Ｐｙｔｈｏｎ),
  - (Ｐｙｔｈｏｎ, ＰＹＴＨＯＮ)
- Transliteration
  - (バイオリン,Violin)
- 読みに関連した変異
  - (100メートル, 100m), (10分, 十分)
  - (憂鬱, 憂うつ), (憂うつ, ユーウツ)
  - (肩かけ, 肩掛け), (問い合せ, 問合せ)
  - (ヴァイオリン, バイオリン)
  - (オーソリティー, オーソリティ)
- 複合語
  - (政府・日銀, 政府日銀)
  - (PHP-MySQL, PHP MySQL)
- 順序の変異
  - (製品・技術, 技術・製品)
- 上記の組合わせ
  - (百メートル, 100m),
  - (海へび, ウミヘビ),
  - (Sカーブ, Ｓ字curve)

異表記とは何(であるべき)か? 3/3

- 風間ら (2009) に文脈類似度データに基づいた事前調査から判明したこと
  - Webデータを相手にすると,
    - 誤表記と異表記との境界
    - 誤用と異表記との境界
    - 新標記と異表記との境界
  - を明確にする必要が生じる
- これらは従来の異表記の定義では問題にされていない

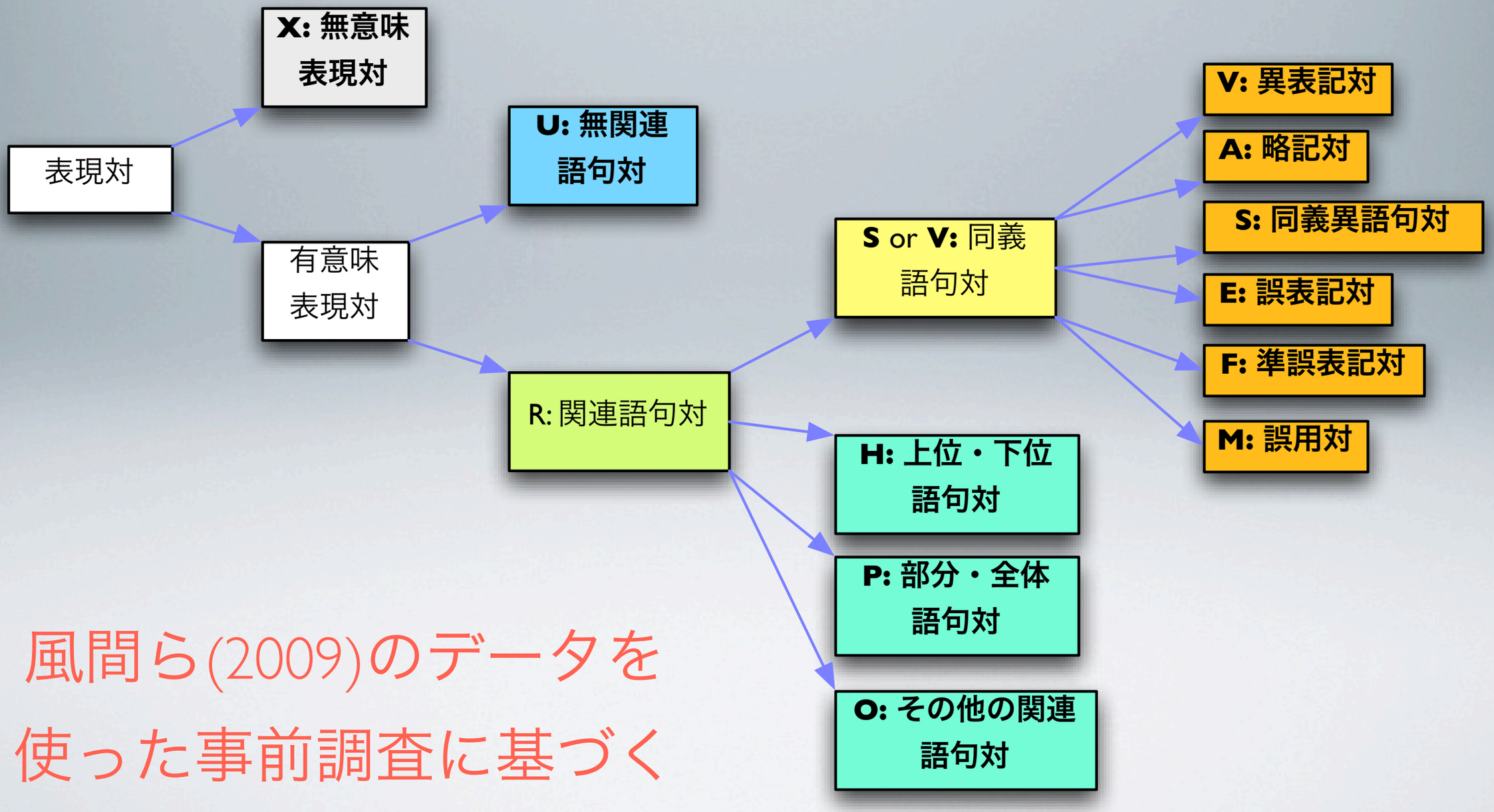
異表記と非異表記の境界 1/2

- 誤表記 (と思しき表記) が係わる対
  - a. (ウェートレス, ウエトレス), b. (ウェートレス, ウェトレス), c. (ウェイトレス, ウェトレス)
- 誤用 (と思しき使用) が係わる対
  - a. (精算金, 清算金), b. (化学兵器, 科学兵器)
- 省略表記が係わる対 1
  - a. (早稲田大学, 早大), b. (医科大学生, 医大生), c. (早稲田大学, 早稲田大), d. (医科大学生, 医科大生) e. (早稲田大, 早稲田)
- 省略表記が係わる対 2
  - a. (ハンセン病患者, ハンセン病者), b. (S 字カーブ, S カーブ), c. (土曜・日曜, 土・日曜), d. (土曜日・日曜日, 土・日曜日), e. (土曜日・日曜日, 土曜・日曜日)

異表記と非異表記の境界 2/2

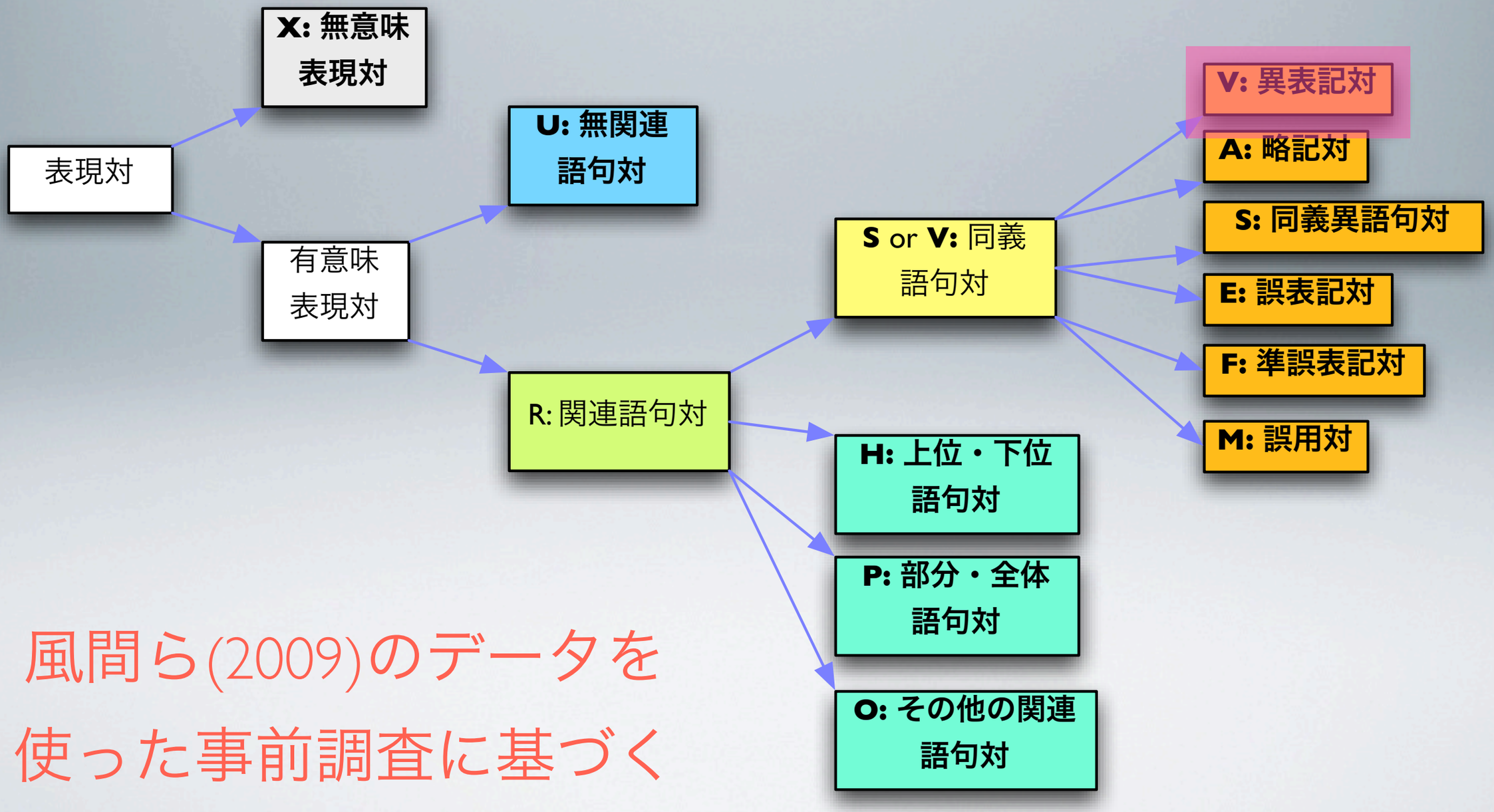
- 同語異表記対と同義異語対の区別が困難な場合がある
- 例
  - (問い合わせ, お問い合わせ), (S字カーブ, Sカーブ), (佐藤, 佐藤さん), (慶応, 慶応大) [addition/deletion]
  - (慶応大学, 慶大), (医科大学, 医大), (工科大学, ??工大), (工業大学, ?工大) [abbreviation]
  - (慶応大学, 慶応大), (大国主命, 大国主) [incomplete abbreviation?]
- 対処
  - 排他分類は不可能だと割り切り，これらは同語異表記対と同義異語対の境界例とするしかない
  - どちらと見なすのがよいかは用途による (application-dependent)

語句対の分類体系 (簡略版)



風間ら(2009)のデータを
使った事前調査に基づく

語句対の分類体系 (簡略版)



風間ら(2009)のデータを
使った事前調査に基づく

提案する異表記対の定義 I/4

- 同義語対と異表記対を意識的に区別している理由
  - 同語異表記対は (定義により) 同一語の異なる標記の対で,
  - 同義異語対は (定義により) 同一の対象を指示する異なる語の対
  - 元の語形を特定できるとは限らないため, 標記の変異 (notational variants) という用語は避けた
- 問題: 同語性の操作的な定義はない
  - それを明示するのは現状では無理
  - “定義することはできないが, 見たらそれとわかる” ような対象の例

提案する異表記対の定義 2/4

- 次の条件を満足する文字列 *s* と *t* の対は異表記対 (allographic pair) である:

A. *s* と *t* が同一でない文字列である (異形条件).

B. *s* と *t* が同一の語を表わす (同一語条件),

- 注意:
  - A条件は自動認識可能だが, 条件Bには人の判断が必要
  - ヒトの評定はどれぐらい信頼できるか? (Fleiss' κ=0.8113 [*n*=3])

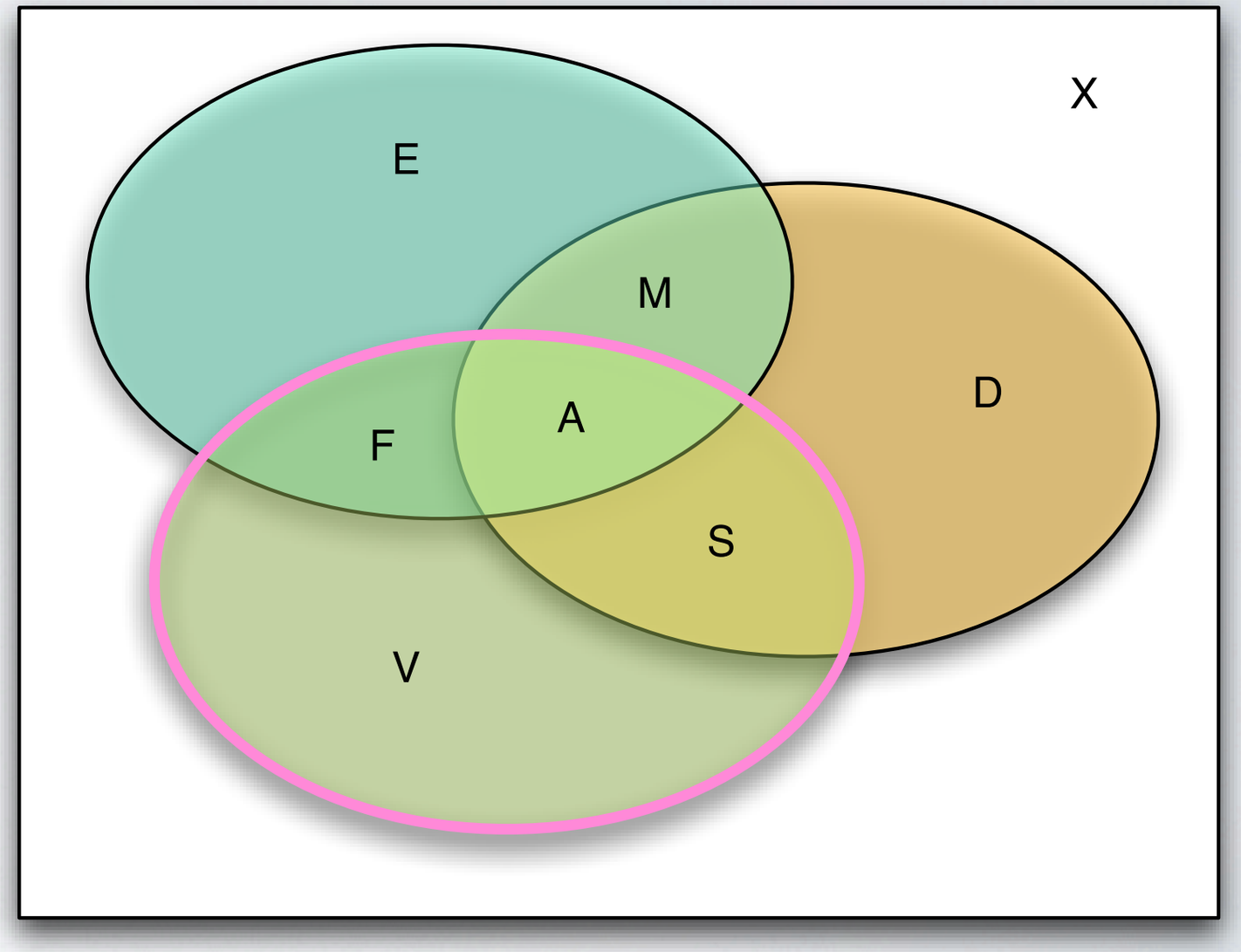
評定の一致度 (Fleiss' kappa)

- Task 1
  - Sample = 3000/15271
  - # raters =3
- Task 2
  - Sample = 3000/16001
  - # raters = 3
- Remark
  - {e, m, u} の区別で作業者間で不一致があった

Labels	**v**, **w**, **s**, **r**, **o**, **e**, **m**, **u**	**v**, **w**, **s**, **r**, **o**	**v**, **w**, **r** (=**s**), **o**, **m**	**v**, **w**, **r** (=**s**), **o**	**v**, **w**, **o** (=**r**,**s**,**e**, **m**,**u**)
Task 1	NA	NA	0.7576	0.7649	0.8113
Task 2	0.5936	0.6040	0.61659	0.6134	0.7536

提案する異表記対の定義 3/4

- 同義対 α = {V, S, F, A}, 異語対 β = {D, S, M, A}, 異形対 γ = {E, F, M, A}
- 単純類
  - **V**: 異表記対, **E**: 誤表記対, **D**: 異語対
- 複合類
  - **S**: 同義異語対, **F**: 準誤表記対,
  - **M**: 誤用対, **A**: 略記対



提案する異表記対の定義 4/4

- α. 同義な対:
  - w1 と w2 とが同義な語句の対であるなら, w1 と w2 は α の要素
- β. 異語の対:
  - w1 と w2 とが (意味の異同は問題にしな いで) 異なる語の対であるなら, それらは β の要素
- γ. 正式形/異形の区別をもつ対:
  - w1 と w2 の一方 が正式な語 (形) であり, 他方が非正式な語 (形) ならば, γ の要素 (ただし誤表記は非正式な表記の特殊例).

異表記対と境界例の関係 2/2

- V (variants) の要素となるのは,一方が他方の同語異表記対の場合. 例は
  - (餃子,ギョウザ) や (ギョウザ, ぎょうざ).
- S (synonyms) の要素となるのは w1, w2 が同義異語対の場合. 例は
  - (大学闘争, 学園闘争), (単独首位, 単独トップ)
- D (distincts) の要素となるのは,w1 と w2 が二つの異語であり, かつ異義語の場合である
  - D には関連語対と無関連語対のすべてが含まれる.
- A (acronymic pairs) の要素となるのは, 一方が正式形で他方がその省略形の場合. 例は
  - (早稲田大学, 早大), (短期大学, 短大)
- M (misuses) の要素となるのは, w1, w2 が異義語だが, 時に一方が他方の意味で誤用される場合. 例は
  - (化学兵器, 科学兵器), (清算, 精算)
- E (errors) の要素となるのは, 一方が用法が確認できない誤記の場合. 例は
  - (思い出, い出)
- F (faulties) の要素となるのは, 誤表記と見なされるべき表記が正表記と同義になる場合. 例は
  - (サンドバッグ, サンドバック), (シミュレーション, シュミレーション)
- 補集合 X (extra) の要素となるのは,対の両方が有意味な語句でない文字列の対である例は
  - (らい手, たい手)

厄介な例の帰属場所 (暫定的)

- ＥとＦの境界例=必要に応じて認識してよい
  - (ウェイトレス, ウェトレス)
- ＶとＳの境界例=必要に応じて異表記と認識してよい
  - (早稲田大, 早稲田)
- ＶとＳかＶとＦの境界例=必要に応じて認識してよい
  - (ハンセン病患者, ハンセン病者)
- ただ，これらの配置が適切かどうかは疑問の余地がある.
  - AとF，AとS の境界例も存在するはず

VとSかVとFの境界例

- 非拘束名簿< 方 >式
- 応急処置< 方 >法
- 元衆< 議 >院議員
- 低減< 対 >策
- 自民< 党 >議員
- カード利用< 金 >額
- 女< 性 >騎士
- 米< 国 >軍人
- 福岡< 市 >近郊
- 食品医薬< 品 >局
- 産業< 用 >機器
- 専門学< 校 >生
- アユタヤ< 王 >朝
- 休息< 場 >所
- 脂肪含< 有 >量
- 排< 気 >ガス中
- 高齢< 者 >福祉課
- 食< 料 >品製造業
- 土産< 物 >店
- つなぎ< 目 >部分
- 老朽< 化 >施設
- 中国< 人 >選手
- 製造< 経 >費
- 文< 房 >具屋

A<B>C は部分文字列Bの脱落可能性を表わす

“w1 + w2”で，w1の末の一文字か，w2の頭の一文字が脱落するタイプは生産的で，これが異表記対でないと相当の取りこぼしが生じる恐れあり

今後の課題

- 体現以外の語句対への対処
  - 風間ら(2009)は，文脈類似度の高い名詞句のデータ
- 非対称性の導入
  - (A, B) を非対称な含意関係だと見なす
    - (半紙, はんし), (藩士, はんし) ☜A=>Bの含意成立で，右が左の異表記
    - (はんし, 半紙), (はんし, 藩士) ☜A=>Bの含意不成立で，右が左の異表記とは言えない
  - 簡単な対処
    - 標記対 (A, B) に関して，Aが B より低頻度なら，(A, B) は異表記対と見なして良いが，(B, A) はそうではない
    - これをやれば評価が2倍にはならない

まとめ

- 異表記とは何(であるべき)か?
  - 日本語異表記の複雑性
    - (1) 異表記対と誤表記対，(2) 異表記と編集距離の近い同義語対の区別が難しい場合があることを指摘
  - その問題を解決するための提案
    - (1) 同義条件 (2) 異語条件 (3) 標記の正式性条件の三つを組み合わせた異表記の定義を提唱した
    - 用途に応じて異表記の認識範囲を変更できる
- 異表記データを ALAGIN フォーラムから公開予定
  - 正例約3万，負例約7万

謝辞

- 次の方(々)から有益な意見を頂きました．この場を借りて感謝
  - 藤田 篤 (はこだて未来大学)

Thank you for your Attention

付録: 大規模異表記対データの構築

異表記対の分類器

- 小島ら (2010) [A4-1] が異表記対のSVM分類器を開発
  - F値 = 90%を達成し，それなりに高性能
  - 教師データ
    - 後述の Task 1 の結果の {**v**} を正例，{**m**, **r**, **o**} を負例
- 言語資源
  - 文脈類似語 (風間ら2009) のうち，編集距離が近い語句の対を SVM で分類し，上位20万対を人手評価
  - ALAGIN Forum (http://www.alagin.jp) を通じて公開予定

分類器の出力サンプル

SVM Score	Candidate	Eval
1.443460	トレジャー < ・ > ハンター	1
1.328230	ウ < ィ | イ > ダーｉｎゼリー	1
1.026730	お手当 < て >	1
0.938527	メ < イ | ー > キャップ	1
0.478044	プ < ク | ス > ッ	0
0.429747	< 第 > ４コーナー	1
0.303877	公 < 的 > 企業	1
0.302481	元町・中華街 < 駅 >	0
0.213331	１２世紀 < 末 >	0
0.019915	< 約 > ６ｍｍ	0

Score	Candidate	Eval
-0.201110	< 俺 | ー > 明日	0
-0.213039	< 売 | と > らないこと	0
-0.224810	十 < 二 > 時半	0
-0.296413	１ < 階 > 天井	0
-0.330489	< 回 | と > っていたこと	0
-0.407557	３００ < ｋ | Ｋ > ｍ	1
-0.414205	< １ > ６時近く	0
-0.415729	< 東 > アジア経済圏	0
-0.436623	< 焚 | 炊 > き方	0.5
-0.439354	< Ｎ | 報 > ステ	0

課題とデータ

- 標準化された編集距離 *r* が近い語句の対 [*r* = # edits/# chars)]
  - Task 1: 評定者3名が文脈類似語 (風間ら2009) による選別ありデータを評定
    - 10,494 対 (3%) (r = 0.200), 4,777 (3%) pairs (r = 0.167)
  - **Task 2:** 評定者3名が類義性による選別なしデータを評定
    - 11,750 対 (r = 0.200), 4,253 pairs (r = 0.167)
- Tasks 1, 2 の比較により，類似性選別の効果がわかる

教師データに必要な特性

- 文脈類似度が高く，編集距離が近い語句の対が異表記対である確率はそれほど高くないことが学べるようなデータが必要

  - 数字の値の違い
    - a. (1メートル, 2メートル), b. (六大学，七大学)
  - 類義/同義語の偶発的な編集距離の接近
    - a. (用紙トレー, 給紙トレー) b. (大学教官, 大学教員)

- 文脈依存性もある

課題で使われたラベル

- **v**: 異表記対
- **w**: クラスメート (異語類語対)
- **s** 同義異語対
- **r**: 関連語対
- **u**: 無関連語対
- **e, m**: 誤表記対, 誤用対
- **o**: 他の語対

Label	Task 1	Task 2
v	Used	Used
w	Used	Used
s	Unused	Used
r	Used	Used
e	Unused	Used
m	Used	Used
u	Unused	Used
o	Used	Used

クラスの例

- **v**:
  - (１ヶ月程度, １か月程度), (２２６事件, ２・２６事件)
- **w**:
  - (エドワード１世, エドワード６世), (３キロメートル, ８キロメートル)
- **s** [Task 2 only]:
  - (ハンセン病患者, ハンセン病者), (ゴミ処理場, ゴミ処分場)
- **r**:
  - (照度アップ, 強度アップ), (入園申込書, 入所申込書)
- **o**:
  - (返すくらい, 流すくらい), (場そのもの, 顔そのもの)
- **m** or **e** [Task 2 only]:
  - (のキャラクター, 某キャラクター), (ホームページ, 家ホームページ)
- **u** [Task 2 only]: (ラーメン店, ラーメン作), (アーカイヴ, アーカイア)

評定結果 (Task1)

三人評定者のうちの一人が与えた値のカウント

label	r0200	r0167	TOTAL	r0167*
v	627 [6%]	289 [6%]	916 [6%]	634.9 [6%]
w	5878 [56%]	2326 [49%]	8204 [54%]	5109.7 [48%]
m	208 [2%]	163 [3%]	371 [2%]	358.1 [3%]
r	744 [7%]	360 [8%]	1104 [7%]	790.8 [8%]
o	3037 [29%]	1639 [34%]	4676 [31%]	3600.5 [34%]
TOTAL	10494	4777	15271	4777*10494/4777

- **v** の率は 6%ほど
  - *r*=.200 と *r*=.167で違いなし
- **w**の率が *r*=.167で低下
- **m**, **o** の率は *r*=.167で上昇
- **r** の率は *r*=.200 と *r*=.167で変わらず

評定結果 (Task 2)

三人評定者のうちの一人が与えた値のカウント

label	r0200	r0167	TOTAL	r0167*
v	355[2.2%]	156[2.6%]	511[2.3%]	418.2[2.6%]
w	229[13.9%]	907[15.1%]	3136[14.3%]	2431.7[15.1%]
s	201[1.3%]	69[1.2%]	270[1.2%]	185.0[1.2%]
r	3847[24.0%]	1379[23.0%]	5226[23.7%]	3697.1[23.0%]
o	8451[53.0%]	3186[53.2%]	11637[52.8%]	8541.8[53.2%]
e	477[3.0%]	90[1.5%]	567[2.6%]	241.3[1.5%]
m	22[0.1%]	11[0.2%]	33[0.1%]	29.5[0.2%]
u	472[2.9%]	190 [3.2%]	662[3.0%]	509.4[3.2%]
Total	16054	5988	22042	4777*10494/4777

- **v**'s の獲得率は 2%程度
  - r=.200 と r=.167 の場合で違いなし
- データの大半が **o**, それに次いで **r** が多い
- **e** と **m** (と **u**) の区別はほとんど意味なし

比較

label	Task1	Task2	Task2*
v	916[6.0%]	511[2.3%]	354.0[2.3%]
w	8204[53.7%]	3136[14.2%]	2172.7[14.2%]
r(,s)	1104[7.2%]	5496[24.9%]	3807.7[24.9%]
o	4676[30.6%]	11637[52.8%]	8062.3[52.8%]
m(,e,u)	371[2.4%]	1262[5.7%]	874.3[5.7%]
Total	15271	22042	22042*15271/22042

- クラスタリングは次に効果あり:
  - 異表記対 (**v**)
  - 同類語対 (**w**)
- 編集距離が小さい文脈類似語対でも，大半が非関連語対
  - 意外な結果
- 5%ほどのノイズがある
  - 元データの成語性チェック作業の必要性

一致率 (Fleiss' Kappa)

- Task 1
  - Sample = 3000/15271
  - # raters =3
- Task 2
  - Sample = 3000/16001
  - # raters = 3
- Remark
  - {e, m, u} の区別で作業者間で不一致があった

Labels	**v**, **w**, **s**, **r**, **o**, **e**, **m**, **u**	**v**, **w**, **s**, **r**, **o**	**v**, **w**, **r** (=**s**), **o**, **m**	**v**, **w**, **r** (=**s**), **o**	**v**, **w**, **o** (=**r**,**s**,**e**, **m**,**u**)
Task 1	NA	NA	0.7576	0.7649	0.8113
Task 2	0.5936	0.6040	0.61659	0.6134	0.7536

評定者間のズレ

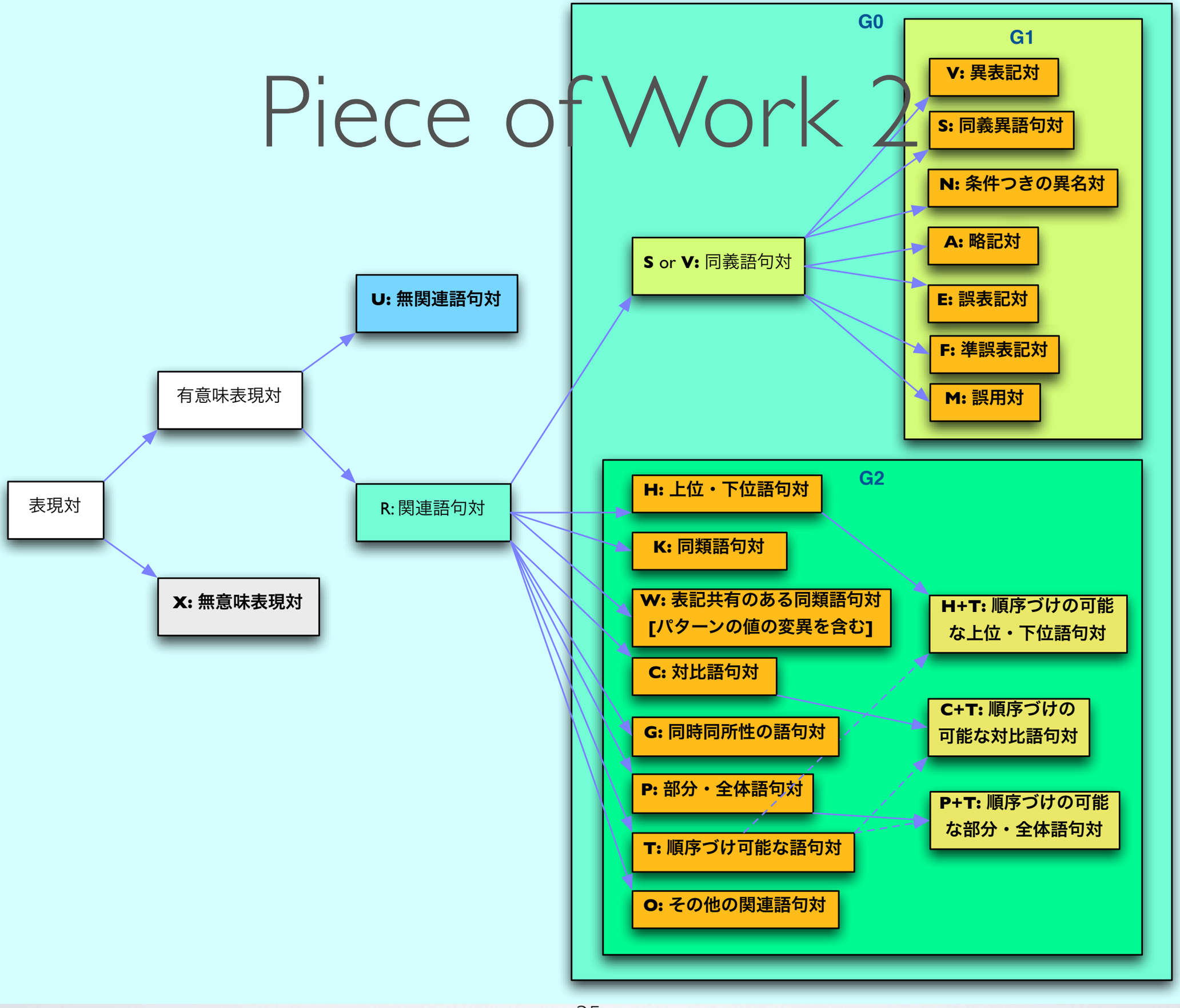
- C の評定が A, B の評定と明らかにズレている
- あるいは A, B の一致率が理不尽に高い??

pair\value	v, w, s, r, o, e, m, u	v, w, o
A, B	0.7697	0.8561
B, C	0.5080	0.6630
B, C	0.4985	0.6889

まとめ

- 文脈類似度による効果があった
  - だが，それでも十分とは言えない
  - 異表記対にとって編集距離が近いことは必要でも十分でもない，それどころか，両者の相関もそれほど高くない
- 基準の有効性
  - 評定結果の一致率は高く，タグは有効だった
  - 比較はしていないので，ちゃんとした評価にはなっていない

Piece of Work 2



Piece of Work 3