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
{ kuroda,kazama,murata,torisawa }@nict.go.jp
(独) 情報通信研究機構 (NICT) MASTAR プロジェクト 言語基盤グループ

1 はじめに

インターネットから自動取得した大規模な言語データを利用して，ユーザーに様々なサービスを提供するのは言語処理技術の応用の一つであり，その一つとして，ネット検索が広く普及している．その一方で処理の高度化によってサービスの高品質化の早期実現が強く望まれている．障害となる問題は幾つか存在するが，その一つが日本語に特有の異表記 (別名「表記の揺れ」) を含めた同義性認識の問題である．

日本語は少なくとも (i) ひらがな，(ii) カタカナ，(iii) 新字体の漢字，(iv) 旧字体の漢字，(v) 全角のアルファベットと (vi) 半角のアルファベット，(vii) 語の境界を表わすための特殊文字 (e.g.,「　」や「・」) を語句内に混在させることを許す．表記の変異は一定の範囲内にあるが，組合わせによって，語句レベルの表現は多様なものになる．

表層情報しか使わない言語処理では文字列の同一性によって語句の同一性を評価するしかない．このため，次の 5 つの表記は文字列としての同一性をもたず，無関係な 5 語句として扱われる:

(1) { ぎょうざ, ぎょーざ, ギョウザ, ギョーザ, 餃子 }

これは検索利用者の直観とズレている．ユーザーが (1) のどれかを入力して検索をかける場合，その意図は (1) のいずれかの表記が現われているページ全体が検索され，結果が自分にとって有用な順に並ぶことである．異表記性が認識されない状態では，入力表記に正確に一致する文字列が現われているページのみに基づく検索結果となり，取りこぼしが生じる．

取りこぼしをなくすためには，(1) にある五つの表記が**同一の語句の異表記**だという知識があればよいが，その知識はまだ十分に体系的な形では存在していない—すでに実用的な異表記処理は行なわれている [2, 1] が，それでも (2)–(5) で取り上げる例を適切に扱えるようなレベルでの処理は行なわれていない．それを構築するための基礎データを作ることが，今回の作業の目的の一つである．[4] は，本稿が提案する異表記認定基準に基づいて作成された正例と負例を使い，高性能な SVM 分類器を作成した．

2 異表記認識への要件

2.1 異表記認定の (見かけ以上の) 難しさ

異表記の認識は難しい課題ではなく，簡単に規則化できると思われるかも知れない．それは対象をどの範囲に決めるかによる．権威ある書き手によって執筆編集され，誤記や誤用をほとんど含まない正式度の高い文章についてそれは真かも知れないが，誤記や誤用を多く含む Web 文書についてそれは必ずしも真ではない．実際，次に挙げる文字列の対 (w1, w2) が異表記かどうかは，精度と被覆率に関するトレードオフを考慮しないで決定できることではない:

(2) 誤表記 (と思しき表記) が係わる対
  a. (ウェイトレス, ウェートレス)
  b. (ウェートレス, ウエトレス), (ウェートレス, ウェトレス), (ウェイトレス, ウェトレス)

(3) 誤用 (と思しき例) が係わる対
  a. (精算金, 清算金)
  b. (化学兵器, 科学兵器)

(4) 省略表記が係わる対 1
  a. (早稲田大学, 早大), (医科大学生, 医大生)
  b. (早稲田大学, 早稲田大), (医科大学生, 医科大生)
  c. (早稲田大, 早稲田)

(5) 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), (土曜日・日曜日, 土・日曜日), (土曜日・日曜日, 土曜・日曜日)

2.1.1 誤表記対と異表記対の区別

(2a) の語句の対は明らかに異表記対だが，(2b) は言語学の伝統的な定義に従えば異表記の対というより，誤表記 (i.e. ウエトレス, ウェトレス) との対だからである．(2b) を異表記と認識するには，異表記の定義を拡張する必要がある．解決案は §3.3 で示す．

2.1.2 略記と異表記との区別

(4) では別の問題が生じている．(4b) を異表記として認識するのは (後述の理由から) 不可能ではないが，(4a) と (4c) の場合はどうか?

(4a) を異表記対と見なすと，次の問題が生じる．第一に，略語一般を異表記として扱うのは，異表記認定のための規則が (機械学習で実装するには) 複雑になりすぎる恐れがある．第二に，異表記認識と同義性認識は概念的に別の課題として区別されるべきだが，それが混同される恐れが生じる．

誤表記の場合であれ，略語の場合であれ，認識したいのは同義性 (正確には指示される対象の同値性) である．異表記認識は同義対認識の一例であるが，特殊な場合でしかないので，同義対を無理に異表記に含める必要はない．私たちは異表記対と同義語対が次の点で異なると考え，その上で異表記対の認識と別に同義性の認識があると考えた:[1)]

(6) a. 異表記対とは，**同一の語**の異なる表記の対である
  b. 同義語対とは，同一の指示対象をもつ語の対である．
  c. 同義異語対とは，同一の指示対象をもつ (ことがある) **異なる語**の対である．

[1)] 語句対の同一性の判断は，ヒトが直観に従って決めるしかないので，これが操作的定義になっていなくても，それ以上のことは望めない．

同義語対は異表記対を含むが，同義異語対は異表記対を含まない．

(4a) のような略式表記と正式表記の等価性を認識する課題は，語句の同義性の認識の問題であり，それが異表記認識の範囲に収まる必要性はない[2]．同義語対と異表記対を概念的に区別しないと，(7) の例も異表記対となり，不自然である:

(7) 異表記ではない同義語対
  a. (用紙トレー, 給紙トレー)
  b. (大学教官, 大学教員)

これらが編集距離が近く，文字列としての類似性が高い同義語対であっても異表記対ではないのは，「用紙」と「給紙」や「教官」と「教員」が (同義になることはあっても) おのおの異なる語だからである．

2.1.3 対称性を前提にしない異表記対認識

伝統的な異表記の定義では，表記対 (w1, w2) が異表記対であるならば，向きを逆にした (w2, w1) も異表記対である．つまり，伝統的な定義では異表記対は対称対である．これは多くの異表記対に関して成立する条件だが，同義性認識との境界例では対称性が満足されない例が出てくる．それが，(4c) の例と (5) に挙げた例である．(4c) の例であれば，「早稲田大」に「早稲田」を代用するのは検索範囲を広げる効果がある．だが「早稲田」に「早稲田大」を代用するのは検索範囲を狭める効果をもつ (「早稲田」は「早稲田駅」「早稲予備校」「早稲田幼稚園」などの略記にもなりえる)．異表記認識，同義性認識の重要な要件が検索式の拡張であれば，[早稲田大 ⇒ 早稲田] という置換は有用だが，[早稲田 ⇒ 早稲田大] という置換は (曖昧性の解消が目的でない限りは) 無用である．

これは (早稲田大, 早稲田) を (単なる同義語対ではなく) 異表記対として，対称性を前提にしない異表記対認識が必要だということである．この基準では，(早稲田大, 早稲田) は異表記対だが，(早稲田, 早稲田大) はそうではない．

(5) の例でも同じことが言える．[ハンセン病患者 ⇒ ハンセン病者] や [S 字カーブ ⇒S カーブ] の置換は成立するだが，その逆の [ハンセン病者 ⇒ ハンセン病患者] や [S カーブ ⇒S 字カーブ] が同義性を保存するかは評価が難しい．また，[土曜日・日曜日 ⇒ 土曜・日曜日] や [土曜日・日曜日 ⇒ 土・日曜日] は，略語の問題と同様に認識のための条件が複雑すぎる可能性がある．

ここでの考察から，異表記について強い定義と弱い定義が可能であることがわかる:

(8) a. 強い定義では，異表記対は対称な対である．
  b. 弱い定義では，異表記対は非対称な対である．

いずれが異表記の定義として有効であるかは利用条件によると考えた方がよいだろう[3]．ただし，調査の結果から見て，すべての場合で非対称性を考慮に入れる必要はなく，それを考慮する必要があるのは (5) のような要素の省略 (か付加) が関与する場合や，漢字の読みを与える場合などに限られるように思われる．

[2] ただし異表記対と同義対は排他的ではなく，(4b) のような例は略語対，かつ異表記対である．

[3] 小島ら [4] の教師データの作成に当っては，弱い条件で対 (w1, w2) の異表記性を認定した．これは判定で方向性を考慮していないことに等しい．この条件の下で，小島らの SVM 分類器は正例 (w1, w2) の逆 (w2, w1) も正例と見なしたデータで訓練された．

同義性の認識にも強い定義と弱い定義がある．対称性を考慮に入れた場合でも，w2 が常に w1 と同義になることを要求するならば，それは同義性の強い定義に基づく判定である．これに対し，w2 が w1 と同義になる場合があれば同義と見なすのであれば，それは同義性の弱い定義に基づく判定である．

3 効果的な異表記認識のための体系

3.1 一般化のために使用したデータ

日本語の異表記が多様であることを考えると，規則化のためのサンプリングの規模が充分に大きいことが不可欠になる．思いつきベースで異表記の類型化を行なうと，得られた一般化で被覆率が不足する危険性が高い．これを回避するため，私たちは [3] の文脈類似語データからサンプルを生成し，類型化に使った．その際，類似度が高い名詞句の対の，標準化された編集距離が小さいものをランダムにサンプリングして一般化のためのデータを得た．

3.2 異表記認識と同義性認識を含む関係認識の構造

もっとも一般的な形として表現の任意の対の関係を評価するという課題を考えると，異表記対の認識と同義語対の認識がどんな課題か明確になる．図 1 が示すように，異表記対 (図 1 の [V]) の認識作業は，同義語句対 (図 1 の [S or V]) の認識作業の特殊な場合であり，同義語句対の認識作業は関連語対 (図 1 の [R]) の認識の特殊な場合である．

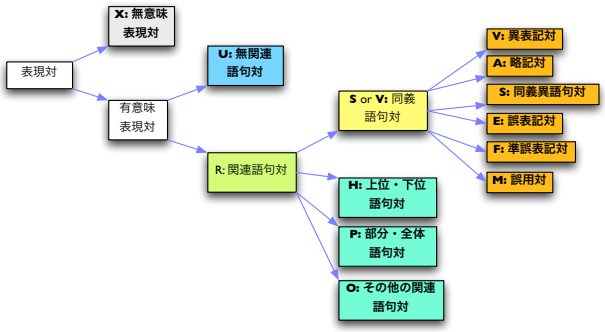

図 1 表現対の分類の一般体系: 関連語句対の下位クラスは略式

3.3 異表記対の体系化

以上の問題点を考慮に入れて，本稿では (i) 同義性 (ii) 異語性 (iii) 表記の変異可能性の三つの条件を組み合わせた異表記認定のための基準を提案する．それに基づくと，(6a) と (6c) の定義を想定した上で，図 2 の $\alpha = \{ V, S, F, A \}$ と $\beta = \{ D, S, M, A \}$ と $\gamma = \{ E, F, M, A \}$ の三つの集合で，異表記対と他の類似クラス (e.g., 同義語対, 誤用対, 誤表記対) との関係をうまく説明できるようになる:

(9) 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/非正式表記の区別をもつ対**: 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な場合とする)．

α, β, γ の重なりを図 2 に示した．これらの集合で定義される様々な部分集合 (V, S, . . .) は以下のように，同義語対の下位分類をうまく記述する:

(10) w1 と w2 の対が
- a. 集合 V の要素となるのは，一方が他方の同語異表記対の場合である．例は (餃子，ギョウザ) や (ギョウザ, ギョーザ).
- b. 集合 S (synonyms) の要素となるのは，w1 と w2 とが同義異語対の場合である．例は (大学闘争, 学園闘争)，(単独首位, 単独トップ).
- c. 集合 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．
- d. 集合 A (acronymic pairs) の要素となるのは，一方が正式形で，他方のその省略形である場合である．例は (早稲田大学, 早大), (短期大学, 短大).
- e. 集合 M (misuses) の要素となるのは，w1 と w2 とが異義語だが，時に一方が他方の意味で誤用される場合である．例は (化学兵器, 科学兵器), (清算, 精算).
- f. 集合 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である．例は (思い出, い出).
- g. 集合 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っている対である．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) など.
- h. 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) など.

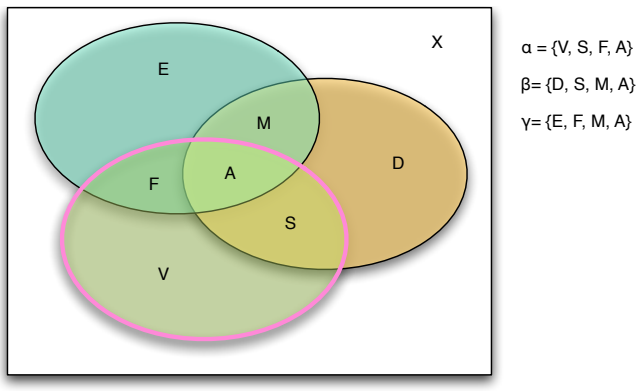

図 2　同義対, 異語対, 正式表記/非正式表記対の関係

3.4　厄介な例の扱い

(2)–(5) で挙げた厄介な例の扱いは次のようになるだろう:

(11)
- a. (ウェイトレス, ウェトレス) のような対は，E と F の境界線上にあるので，異表記と見なす必要はないが同義語対として認識してもよい.
- b. (ハンセン病患者, ハンセン病者) のような対は，F と V の境界線上にあるので，必要に応じて異表記と見なせる.
- c. (早稲田大, 早稲田) のような対は，V と S の境界線上にあるので，必要に応じて異表記と見なせる (ただし方向性の考慮が必要).

4　異表記対の実例と類例の解説

4.1　異表記対 [V] の典型事例集

(6a) の定義に合致する事例は数多くあり，幾つかの下位類が存在する．(12) に下位類と幾つかの例を示す:

(12)
- a. 数字や単位の異表記
  - i. (一リーグ制, １リーグ制)
  - ii. (１００メートル, １００m), (５７ｋｇ, ５７キロ), (５７ｋm, ５７キロ)
  - iii. (３－０, ３対０)
- b. 主に外来語の音の転記の変異に由来する (主にカタカナの) 表記の変異
  - i. (コクピット, コックピット)
  - ii. (ハンナ・アーレント, ハンナ・アレント)
  - iii. (オーソリティ, オーソリティー)
  - iv. (ヴァイオリン, バイオリン)
- c. 字種の変異
  - i. (憂鬱, ゆううつ), (ユーウツ, 憂鬱)
  - ii. (肩掛け, 肩かけ), (お猪口, おちょこ)
  - iii. (辺り, あたり)
  - iv. (当たり, アタリ), (あたり, アタリ)
  - v. (ヘビ, 蛇), (桃, モモ), (ハモ, 鱧)
  - vi. (チリトリ, チリ取り), (竿竹, サオ竹)
  - vii. (長め, 長目)
- d. 外国語の音転記 (transliteration) と元語句との対[4)]
  - i. (オリコンスタイル, ｏｒｉｃｏｎ　ｓｔｙｌｅ)
  - ii. (ATARI, アタリ)
- e. 大文字と小文字の変異
  - i. (Ｋｅｒｎｅｌ, ｋｅｒｎｅｌ)
  - ii. (ｇｒａｐｈ, ＧＲＡＰＨ)
- f. 全角文字と半角文字の変異
  - i. (Ｋｅｒｎｅｌ, Kernel)
  - ii. (GRAPH, ＧＲＡＰＨ)
- g. 空白の有無
  - i. (ＰＨＰＭｙＳＱＬ, ＰＨＰ　ＭｙＳＱＬ)
  - ii. (PHPMySQL, PHP MySQL)
- h. 字体の変異
  - i. (仙台, 仙臺), (渡邊, 渡辺)
- i. 送り仮名の有無
  - i. (長め, 長いめ)
  - ii. (お問い合わせメール, お問合せメール), (お問い合わせメール, お問い合せメール)
- j.「意味の軽い」形態素の付加 (あるいは省略)
  - i. (問い合わせ, お問い合わせ)
  - ii. (Ｓ字カーブ, Ｓカーブ), (大国主命, 大国主)
  - iii. (和田秀樹氏, 和田秀樹)
- k.「・」などの記号の有無
  - i. (政府日銀, 政府・日銀), (京都宇治, 京都・宇治)
  - ii. (京都宇治, 京都/宇治)
- l. 順序の交替
  - i. (製品・技術), (技術・製品)

[4)] なお，(apple, リンゴ) のような原語と訳語の対は同義語対であり，異表記対ではない.

m. 上記の場合の組合わせ
  i. (海へび, ウミヘビ), (チリトリ, ちり取り)
  ii. (S カーブ, S字 curve), (57 ｋｇ, ５７キロ),
  iii. (問合わせメール, お問い合わせメール)
  iv. (ATARI 社, アタリ社), (XBox, Ｘボックス)

4.2 同義語対 [S] の下位分類

本論文では詳しく論じないが，同義性認識では，(13) に示す同義語の下位分類を設けると有効である:

(13) a. 同一の対象 (か概念) が異なる観点で記述されていることが明確な場合 (e.g, (用紙トレー, 給紙トレー), (旧 Mac OS, OS 9 以前), (太平洋戦争, 大東亜戦争))
b. 同一の対象 (か概念) が異なる観点で記述されていることが不明確な場合 (e.g., (おじゃん, オシャカ))

(13b) は類義語と重なるが，前者はそうではない.

4.3 異表記対 [V] と略語対 [A] との境界

(4a) に例を挙げた (省) 略語 (形) (acronyms) は同義語対の特殊な場合で，異表記から区別する必要があると判断した: これは，異表記認定の条件を機械学習で実装可能な程度の一般性に留めておく必要があると考えたからである.

だが，次のような中間的な形態が存在するため，話が少しややこしくなる:

(14) a. (早稲田, 早稲田大), (慶応, 慶応大)
b. (日比谷, 日比谷高)

基準を一貫したものにするには，ここの例と上の (4a) の例との区別が必要である.

先に (12) の (Ｓカーブ, Ｓカーブ) や (和田秀樹, 和田秀樹氏) や (大国主, 大国主命) のような場合も異表記対に含め異表記対に含めると説明した．このことから，(15) にあるような例も異表記対の範囲に含めることになる:

(15) a. (佐藤, 佐藤さん), (佐々木, 佐々木氏)
b. (トリュフォー, トリュフォー監督)
c. (遊撃部隊, 遊撃隊)

(4a) の「早大」は「早稲田大学」の短縮形だが，「早稲田大」は「早稲田大学」の「大学」だけが短縮された形というより，「早稲田」に「大」を付加した語形で，(15) に近い形態だという直観がある．これが正しいならば，「～ 大」「～ 高」「～ 中」「～ 小」のような「意味の軽い」形態素は (15) の「～ 氏」や「～ さん」の接尾語の特殊な場合と考えてよい．これに対し，(4a) のような略語のパターンを語彙的に予想するのは困難である.

この違いに基づくと，(14) と (15) の場合を (4a) の場合から区別するのに有用な基準は次の通りである:

(16) 一方の語句 w1 が複数の語 { x1, x2, ..., xn } からなる複合語 x1· x2 ⋯ xn であり，二つ以上の要素について短縮が起こっている語形 w2 と元の w1 との関係は，単純な異表記の関係ではなく，w2 は (w1 の同値表現としての) w1 の短縮形 (acronyms) である.

(4a) に挙げた事例は (16) が該当するので [A] と認識できる.

4.4 誤記と誤用に関係した例外的なクラス

無意味表現対と有意味表現対の境界が不明確な場合があり，誤表記対 [E], 準誤表記対 [F], 誤用対 [M] の区別を設けた.

4.4.1 (非語も含めた) 誤表記との対 [E]

w1 と w2 の一方が正式な語でないものが現われているのは，次のような場合である:

(17) a. (もらい手, らい手)
b. (シミュレーション, シミュュレーショョン)

これは主に入力時やデータ解析の時の誤りに起因する.

4.4.2 準誤表記対 [F]

誤表記対 [E] に関連して，ごく稀れに異表記と誤 (表) 記の区別が曖昧な場合がある．特に (18) のような場合には，誤表記 (e.g., サンドバック, シュミレーション) の方も慣用化しているという厄介な事情がある.

(18) a. (サンドバッグ, サンドバック)
b. (シミュレーション, シュミレーション)
c. (アフェリエイトサイト, アフィリエイトサイト)

これらの場合には，誤表記対 [E] なのか異表記対 [V] なのかを誰もが同意するように判定するのは難しい.

4.4.3 誤用対 [M]

誤表記とはちがって，一方が他方の誤用になる可能性が考えられる誤用対 [M] がある．(19) に幾つか例を挙げる:

(19) a. (精算金, 清算金) [w2 が w1 の意味で使われるのは正確には誤り]
b. (化学兵器, 科学兵器) [w2 が w1 の意味で使われるのは意味の上では誤りではないが，非標準的]

これは準誤表記対 [F] の場合に似ているが，w1 と w2 は異語なので [F] ではない.

5 終わりに

本稿では (a) 誤表記対と異表記対との曖昧性，(b) 異表記対と同義語対との曖昧性に対応できる異表記対の認識基準を提案したが，それは瑣末な細部にこだわったものに見えるかも知れない．だが，効能や示唆がないわけではない．第一に，誤表記対と異表記対の区別，同義異語対と異表記対の正確な区別は，人手分類の一致率の向上に貢献した[5]．また，本稿の分析から，省略は特異な性質をもつものであることがわかり，より広い範囲で略語の同義性を自動認識のためには，独立の処理モジュールが必要になることも示唆される.

参考文献

[1] K. Masuyama and S. Sekine. Automatic construction of katakana expression variation from large corpus. In *The 10th Annual Meeting of the Association for Natural Language Processing*, 2004.

[2] 荒牧 英治, 今井 健, 梶野 正幸, 美代 賢吾, and 大江 和彦. Support Vector Machine を用いた医学用語の表記ゆれ解消. In **言語処理学会第 *14* 回年次大会**, pages 135–138, 2008.

[3] 風間 淳一, S. De Saeger, 鳥澤 健太郎, and 村田 真樹. 係り受けの確率的クラスタリングを用いた大規模類似語リストの作成. In **言語処理学会第 *15* 回年次大会発表論文集**, pages 84–87, 2009.

[4] 小島正裕, 村田 真樹, 風間 淳一, 黒田 航, 藤田 篤, 荒牧 英治, 土田 正明, 渡辺靖彦, and 鳥澤 健太郎. 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. In **言語処理学会第 *16* 回年次大会発表論文集**, 2010.

[5] ただし，それを示す数値データは本稿執筆時点では用意できなかった.