日本語異表記対データベース (Version 1.1)

初版: 2010 年 03 月 25 日
更新: 2010 年 04 月 21 日 (For Version 1.0)
更新: 2010 年 10 月 20 日 (For Version 1.1)

目 次

はじめに

0.1 公開履歴

1. 2010 年 04 月 21 日: Version 1.0 公開 (人手生成データベースの公開)
   (a) 編集距離=1 の異表記対 [v] の数: 48,067
   (b) 編集距離=1 の準異表記対 [d] の数: 10,730
   (c) 編集距離=1 の同義異語対 [s] 数: 2,758 [非異表記対の一クラス]
1. 2010 年 10 月 21 日: Version 1.1 公開 (自動獲得データベースの追加)
   (a) 編集距離=1 の異表記対の自動獲得結果の数: 約 110-150 万

0.2 ファイルの一覧

- 2010 年 04 月 21 日: Version 1.0 公開　(人手生成データベースの公開)
  - allographic-pairs-v1.sjis.tsv.zip [文字コード Shift-JIS (Windows 環境向け)]
  - allographic-pairs-v1.eucj.tsv.zip [文字コード EUC-JP (Unix/Linux 環境向け)]
  - allographic-pairs-v1.utf8.tsv.zip [文字コード UTF-8 (汎用)]
- 2010 年 10 月 21 日: Version 1.1　(自動獲得データベースの追加)

- allographic-pairs-svm.linear.utf8.gz [文字コード UTF-8 (汎用)、SVM の linear kernel で学習した分類器で獲得した結果]
  - allographic-pairs-svm.linear.s1 (約 139 万の異表記対、 Precision 96.6%レベル)
  - allographic-pairs-svm.linear.s2 (約 153 万の異表記対、 Precision 95%レベル)
- allographic-pairs-svm.poly.utf8.gz [文字コード UTF-8 (汎用)、SVM の polynomial kernel (degree 2) で学習した分類器で獲得した結果]
  - allographic-pairs-svm.poly.s1 (約 115 万の異表記対、 Precision 97.4%レベル)
  - allographic-pairs-svm.poly.s2 (約 130 万の異表記対、 Precision 95%レベル)

0.3 利用条件

本データベースの利用には、(独) 情報通信研究機構と利用許諾契約を結ぶ必要があります。詳しくは、

`http://www.alagin.jp`

をご覧ください。

0.4 利用に関する注意

本データベースは、インターネットホームページ等、(独) 情報通信研究機構以外の第三者が作成した文書等のデータから、語彙の抽出及び統計処理等によって作成されたものです。そのため本データベースの内容は、(独) 情報通信研究機構の主体的な意思決定・判断を示すものではあ

りません。本データベースの生成は、電子的な方法又は一様の選別基準による機械的判定によって行われています。そのため本データベースの内容の正確性、真実性及び相当性は一切保証されません。以上の理由により、(独) 情報通信研究機構は、本データベースの内容について、責任を負いかねます。本データベースの使用に関連して生ずる損失、損害等についても、一切責任を負いかねます。本データベースには、意図せず、第三者への誹謗中傷、差別用語、個人情報などが含まれている場合があります。本データベースを利用の際はこれらによる権利侵害に十分な注意をお願いいたします。利用者においては、本データベースの以上の特質をよくご理解の上で、本データベースをご利用下さい。

0.5 本データベースに関する問い合わせ先

独立行政法人情報通信研究機構
知識創成コミュニケーション研究センター
MASTAR プロジェクト 言語基盤グループ
email: alagin-lr@khn.nict.go.jp

以下、セクション 1 ではデータの基本的な特徴を説明し、 セクション 2 と 3 でデータ構造と付与情報及びデータの構築法の詳細を説明します。

1　データの簡単な解説

本データは、文字レベルの編集距離の近い、日本語の語句の異表記対 (あるいは「表記揺れの対」) の正例と負例を集めたデータです。

1.1　人手生成データベース

人手生成データベースは、類似度の高い語句の集合から自動で生成した異表記対候補を人手でチェックしたデータであり、確実な正例として 48,067 例、確実な負例として 2,758 例、正例か負例か確実に判断できない例として 10,730 例を含みます。現時点での負例は正例との区別が特に困難な同義語句対のみです。

異表記対の正確な定義は §2.1.1 に示しますが、簡単に言うと単一の語句が幾つかの異なる表記 $\{f_1$、f_2、…、$f_n\}$ で使われている場合、f_i と f_j $(i \neq j)$ は異表記対です。具体例を挙げると、{ギョウザ、ギョーザ、ぎょうざ、ぎょーざ、餃子} は異表記の集合です。これから得られる対 [ギョウザ、ギョーザ]、[ギョウザ、ぎょうざ]、[ギョウザ、餃子]、[ギョーザ、ぎょうざ]、[ギョーザ、ぎょーざ]、[ギョーザ、餃子] … は異表記対です。

一般的な異表記の定義はこの通りですが、本データで収集の対象にしているのは編集距離が近い異表記対 (原則として編集距離=1 の場合) のみです。編集距離が近いとは、簡単に言うと、「語句の間で異なる文字の数が少ない」という意味です。編集距離の近い異表記対の例は

- “[Ｃｅｎｔｅｒ、ｃｅｎｔｅｒ]” (大文字と小文字の違い)、
- “[ゴミ置き場、ゴミ置場]” (送り仮名の有無の違い)、
- “[ギタープレー、ギタープレイ]” (語末の「ー」と「イ」の違い)、
- “[ツィンマーマン、ツィマーマン]” (「ン」の有無の違い)、

単語 1 (語句 1)< *tab* > 単語 2 (語句 1)< *tab* > スコア
アクションクリエイター < *tab* > アクションクリエーター < *tab* > 1.60129
アクションコーディネイト < *tab* > アクションコーディネート < *tab* > 1.10272
ウォーマシン < *tab* > ウォーマシーン < *tab* >1.21594
世界中がアイ・ラブ・ユー < *tab* > 世界中がアイ・ラヴ・ユー < *tab* >1.09566
世界選手権の成績 < *tab* > 世界選手権成績 < *tab* >0.401025
主なみどころ < *tab* > 主な見どころ < *tab* > 1.19975
主な勝ち鞍 < *tab* > 主な勝鞍 < *tab* >0.24511
仲町通 < *tab* > 仲町通り < *tab* >1.00205

表 1: 自動獲得データベースのフォーマットと例:　スコアは自動獲得するために作った分類器のスコアで、このスコアが高ければ、高信頼度の異表記対結果と考えます。単語と単語、単語とスコアはタブで区切っています。

- “[ブルース・スプリングスティーン、 ブルーススプリングスティーン]” (「・」の有無の違い) などです。

上に挙げた [ギョウザ、 ギョーザ]、 [ギョウザ、 ぎょうざ]、 [ギョウザ、 餃子]、 [ギョーザ、 ぎょうざ]、 [ギョーザ、 ぎょーざ]、 [ギョーザ、餃子] . . . は異表記対ですが、編集距離が遠いので本データには含まれません。

1.2　自動獲得データベース

このデータは Web 1 億ページに出現する頻度上位約 1,000 万の日本語語句（主として単語）に対して「編集距離 1 の異表記対」（今後、簡単に「異表記対」と呼ぶ）を自動獲得した結果です。自動獲得した異表記対の結果は表 1 のような形式を持っています。

自動獲得した異表記対データの量は、使った分類器と異表記対を決める閾値によって 4 種類があり、それぞれ約 110 万以上の異表記対を含みます。

1.3　本データの典型的な使い方

本データは特定の用途に特化したものではありませんが、典型的な用途として次のような利用法が考えられます:

- 被覆率の向上を目的とした検索式 (query) の拡張
- データの標準化 (regularization)

「検索式 (query) の拡張」の例は、ユーザーが検索に「餃子」と入力している時に、その検索条件を「餃子 OR ギョーザ OR ギョウザ OR ぎょうざ OR ぎょーざ」に自動展開する場合です (ただし、本データは「餃子」の場合のような編集距離が遠い異表記対は収録していません)。「データの標準化」は、Web から自動獲得した名詞集合のクラスタリングをしたい時に、見かけの異なり数を減らすという操作です。

2　人手生成データベース

2.1　データの作成手順

本データは黒田ら (2010) の方法に基づいて、次の手順で作成されました:

- 手順

Step 1　風間ら (2009) の手法で構築された名詞句のクラスター化データを基にして、編集距離の近い語句対を異表記対の候補として 20 万対生成する (これにより、編集距離が近く、かつ意味の類似度の高い対を多数獲得できる)。

Step 2　これらの候補を、下の (§2.1.1 と図 1) に示す 10 個の基準 [s、 a、 v、 e、 f、 m、 w、 r、 u、 x] で人手分類する。

Step 3　この段階で小島ら (2010) の開発した分類器を使って人手評定で間違って異表記に分類された可能性のある事例 (false positives) と間違って非異表記に分類された可能性のある事例 (false negatives) の候補を生成し、それらを重点的に検査した。

– 最終チェックで v と s の境界に幅を持たせるため下記の場合、[d] とし判定する。
  * 3 つの基準 [e、 f、 m] で判定される事例
  * かつ [v] と [s] の排他分類が難しい事例

この区別の下で、

- “[v]”は「異表記対の正例」、
- “[d] は「正例か負例か一概には決められない例」、
- その他の場合の [s、 w、 r、 u] が「異表記対の負例」となる ([x] は負例には含めない)。

Step 2 と 3 の作業は、図 1 の分類体系によって行います。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す。

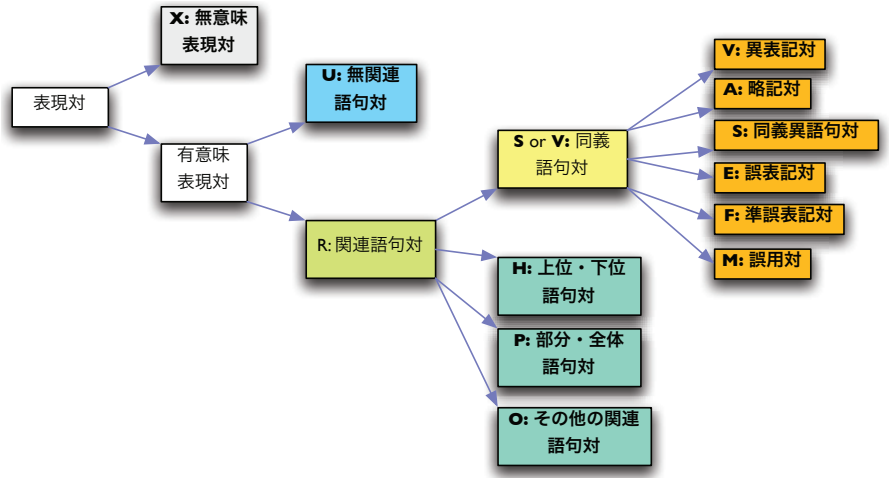

図 1: 分類の一般体系: 関連語句対の下位クラスは略式。10 個の基準の一つ W は A の下位分類と考えます。

2.1.1 人手分類の基準

s: 同義異語句対: 同じ対象を指示する (ことのある) 異なる語句の対である場合。例えば
[用紙トレイ、給紙トレイ]、[学園闘争、学園紛争]、[単独首位、単独トップ]、[パイプ内、配管内]、[ガウス分布、正規分布]、[買い手、売る相手]、[責任逃れ、言い逃れ]

a: 略語対: 同じ語句の異なる表記の対だが、一方が他方の略式表記になっている場合。例えば
[慶応大学、慶大]、[短期大学、短大]、[HDD、ハードディスクドライブ]

w: a[略語対] の特殊な場合、形態素共有のある同類語: 共通の上位語をもつ同類語で (主に語句末で) 形態素を共有する場合。例えば
[中国、韓国]、[二日、三日]、[土曜日、日曜日]

v: 同語異表記対: [a] を除いて同じ語句の異なる表記の対である場合。例えば
[一リーグ制、1リーグ制]、[100メートル、100m]、[57キロ、57k]、[ハンナ・アーレント、ハンナ・アレント]、[オーソリティ、オーソリティー]、[憂鬱、ゆううつ]、[肩掛け、肩かけ]、[アタリ、ATARI]、[Kernel、kernel]、[PHPMySQL、PHP MySQL]、[お問い合わせ、問合せ]、[海ヘビ、うみへび]

e: 誤表記対: v の特殊な場合で、一方が他方の誤表記だと判断できる場合。例えば
[メールアドレス、ルアドレス]、[もらい手、らい手]、[シミュレーション、シミュュレーション]

f: 準誤表記対: 本来は誤記だと思われる表記が正用化していると判断できる場合。例えば

[サンドバッグ、 サンドバック] (cf. バック転 vs *バッグ転)、[シミュレーション、 シュミレーション]

m: 誤用対: s の特殊な場合で、異なる語句が変換ミスなどによって偶発的に同じ意味で使われていると判断できる場合。例えば
[精算金、 清算金]、 [化学兵器、 科学兵器]

r: 関連語対: 同義ではないが、はっきりと認識できる関連性がある場合。はっきり認識できる関連性があるとは、次のような関係が認識できることである:

r-1: 上位語と下位語の対。図 1 の H に対応する。例えば
[柴犬、 犬]、[再婚、 結婚]

r-2: 部分を表わす語句と全体を表わす語句との対。図 1 の P に対応する。例えば
[太平洋戦争、 第二次世界大戦]、[椅子、 背もたれ]、[ジョン・レノン、 ビートルズ]

r-3: 対比語の対 (ただし、これは同類語の特殊な場合)。例えば
[右側、左側]、[高抵抗、 低抵抗]

r-4: 共通の上位語をもつ同類語で、形態素共有のない語句の対。例えば
[タイ、 アルゼンチン]、[イワシ、 サンマ]、

r-5: 共通の全体をもつ語句の対 (これは対比語の特別な場合)。例えば
[ジョン・レノン、ポール・マッカートニー] [リール、 釣竿]、[エンジン、 タイヤ]

r-6: 時間上の順序づけが可能な語句の対。例えば
[離婚、結婚]、 [再婚、 離婚]、[出産、 妊娠]

r-1 から r-6 までの以外にも関連性のタイプは幾つかあるが、すべて

をここで網羅することは適当できない。重要なのは、r[関連語対] が次の u[無関連語対] や x[無意味語対] と区別できるという点である。

u: 無関連語対: 両方の語句が意味をなすが、はっきりと認識できる関連性がない場合。例えば
[風習、 アーム]、[船体、 仙臺]

x: 無意味語対: 少なくとも一方が意味をなさない語句である場合。例えば
[い出、 思い出]、[もら、 もち]

2.2 データの見本

ABD と ACD が (B と C の違いについて) 異表記対であることを表すために、"$A < B \mid C > D$"という表記を用います。その特別な場合として、"$A < B > D$" は、ABD と AD が異表記対であることを表わすことにします。

(1) に異表記対の実例 [v] を、(2) に準異表記対の実例 [d] を、(3) に異語同義対の実例 [s] (異表記の負例) を、おのおの 20 例ほど示します:

1. 編集距離=1 の異表記対の例 (20 対)

   (a) 第＜一｜１＞週目

   (b) ＜４｜四＞カ月後

   (c) Ｆｌａｓｈ　＜Ｐ｜ｐ＞ｌａｙｅｒ

   (d) ＜Ｃ｜ｃ＞ｅｎｔｅｒ

   (e) ＜Ａ｜ａ＞ｇａｉｎ

   (f) ゴミ置＜き＞場

   (g) 割＜り＞引き価格

   (h) ギタープレ＜ー｜イ＞

   (i) ブルース＜・＞スプリングスティーン

(j) ウ < イ | ィ > ルス性

(k) ツィ < ン > マーマン

(l) テディ < ー > ベアー

(m) エロゲ < ー > 板

(n) キョ < ー | ウ > コ

(o) 普及 < ・ > 定着

(p) < ご > 希望どうり

(q) そこ < い > らじゅう

(r) < 龍 | 竜 > 神様

(s) < は | 剥 > がすため

(t) お < 替 | か > わり

2. 編集距離=1 の準異表記対の例 (20 対)

(a) 法 < 律 > 違反

(b) 補足 < 的 > 給付

(c) 調査 < 手 > 法

(d) 株 < 式 > 取得

(e) 米 < 国 > 本社

(f) 手数料 < 金 > 額

(g) 胴体下 < 部 >

(h) 満州 < 国 > 軍

(i) 動作性 < 能 >

(j) 土曜・日曜 < 日 >

(k) 依頼者 < 様 >

(l) シレーナ < 様 >

(m) 森山大道＜氏＞

(n) ネルソン・マンデラ＜氏＞

(o) ドクター中松＜氏＞

(p) 川内康範＜氏＞

(q) 山本夏彦＜氏＞

(r) 府立大＜学＞

(s) 京都産業大＜学＞

(t) 地元新聞＜社＞

3. 編集距離=1 の異語同義対 (非異表記対の一種) の例 (20 対)

(a) ＜社＞日本青年会議所

(b) コンスタンティヌス＜帝＞

(c) インテル＜社＞

(d) シックスアパート＜社＞

(e) ＧｅｏＴｒｕｓｔ＜社＞

(f) Ｋｏｄａｋ＜社＞

(g) 米アップル＜社＞

(h) Ｓｉｅｍｅｎｓ＜社＞

(i) フィナンシャル・タイムズ＜紙＞

(j) ビハール＜州＞

(k) 北海道札幌＜市＞

(l) 差別的＜だ＞

(m) 常識的＜だ＞

(n) エリア＜以＞外

(o) 同日＜以＞後

(p) 車いす < 専 > 用

(q) < 身 > 体各部

(r) 李 < 前 > 総統

(s) 中曽根 < 元 > 首相

(t) 男 < の > 子同士

2.3 注意 1: 準異表記対の扱い

準異表記には、異表記か否かを排他分類するのが困難な場合 (e.g. 手数料 < 金 > 額) の他、誤表記 (e.g.、 オヤジギャ < グ | ク >)、新表記 (e.g.、< へ | 屁 > ロス) なども含まれます。

2.4 注意 2: 非対称性の扱い

現時点では異表記対の非対称性は考慮していません。異表記対は対称対だと考えられることが多いようですが、実際には含意の方向があり、非対称です。実例を挙げると、[お < 替 | か > わり] の対と [お < か | 替 > わり] の対は異なる含意を表わしており、別に扱わなければならないことがわかります。[お替わり ⇒ おかわり] は無条件に真だが、[おかわり ⇒ お替わり] は ([おかわり ⇒ お変わり] や [おかわり ⇒ お代わり] がある以上) 無条件に真ではないからです。精度の向上のためには、この非対称性を考慮に入れたデータを用意すべきなのですが、今の時点ではそこまで対応しておりません。将来の更新で非対称性を考慮したいと考えています。

3 自動獲得データベース

3.1 作成方法

データは「異表記対候補の生成」と「生成された候補の異表記対判定」の 2 つのステップで作成されました。(候補の生成や判定の詳しい説明は

小島ら (2010) をご参照下さい。）

「異表記対候補の生成」と「候補の異表記対判定」の学習データは ALAGIN で公開されている「日本語異表記対データベース（Version 1）」(「人手生成異表記対データベース」）と「基本的意味関係の事例ベース（Version 1）」を元に編集距離が 1 の異表記対を自動的に生成して準備しました。異表記対の学習データは約 40,000 の正例（異表記対である語句ペア）と約 100,000 の負例（異表記対ではない語句ペア）を含んでいます。

3.1.1 異表記対候補の生成

異表記対学習データの正例約 40,000 から約 3,500 の異表記対の生成パターン（文脈独立なパターン）を取り出します。例えば、「アクションコーディネイト、アクションコーディネート」の異表記対からは異表記対の生成パターン「イ→ー」が取り出せます。この異表記対の生成パターンを Web 1 億ページに出現する頻度上位約 1,000 万の日本語語句（主として単語）に適用し、生成された異表記候補が上記 1,000 万語中に見つかれば元の語句と組み合わせて異表記対の候補と考えます。この結果、約 1,200 万の異表記対の候補を生成しました。

3.1.2 候補の異表記対判定

候補の異表記対判定は機械学習によって構築した分類器によって行いました。分類器の学習のため、異表記対学習データの正例と負例が持つ特徴を文字レベル（異なる文字の種類、文字の変換パターン、文字レベル文脈情報など)、形態素レベル（異なる文字を含む形態素及び品詞、形態素レベルの文脈など)、辞書レベル（辞書で同じ意味に属すか、基本形が同じかなど）の素性で表現します。異表記対の候補も同様の素性で表現します。最終的には学習した分類器を異表記対の候補に適用して異表記対であるかを判定します。分類器のために使った素性のリストを表 4 と 5 に挙げます。S1 から S30 までの文字レベルの素性では、異表記対を編集箇所（異表記対候補の違う文字の部分）の文字と文字の種類及びその

編集箇所の前後文字を文字レベル文脈として考えて素性で表現しました。S31 から S34 までは形態素レベルの素性、S45 と S46 は JUMAN の辞書を使った辞書レベルの素性です。S35 から S44 までは異表記対によく表すパターンを素性として表現しました。

3.1.3 異表記対候補判定の評価

分類器は SVMs の 2 つの Kernel（Linear kernel と Polynomial Kernel の degree2）で学習しました。35,021 の評価データ（8,456 の正例と 26,565 の負例を含む）を使って評価した結果を表 2、3 に示します。評価の際は、分類器のスコアが閾値より高ければ「異表記対」と判定し、閾値より低ければ「異表記対ではない」と判定しました。表 2 と 3 は異表記対候補判定の結果得られた異表記対を含むファイルを「allographic-pairs-svm.linear.s1」、「allographic-pairs-svm.linear.s2」、「allographic-pairs-svm.poly.s1」、「allographic-pairs-svm.poly.s2」と表記します。評価の結果、学習に適用した SVMs の Kernel の種類と閾値によって少し差がありますが、全ての設定で 95%以上の F 値が得られることがわかりました。

ファイル名	閾値	Precision	Recall	F-value
allographic-pairs-svm.linear.s1 異表記対の数：約 139 万	0	96.6	95.3	95.9
allographic-pairs-svm.linear.s2 異表記対の数：約 153 万	−0.29	95.0	96.5	95.7

表 2: Linear Kernel の評価結果

参照文献

- 風間 淳一、 デサーガ ステイン、 鳥澤 健太郎 and 村田 真樹 (2009). 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 **言語処理学会第第 15 回年次大会発表論文集**、 pp. 84–87.

ファイル名	閾値	Precision	Recall	F-value
allographic-pairs-svm.poly.s1 (異表記対の数：約 115 万)	0	97.4	94.1	95.7
allographic-pairs-svm.poly.s2 (異表記対の数：約 130 万)	−0.33	95.0	96.2	95.6

表 3: Polynomial Kernel(degree 2) の評価結果

- 黒田 航、 風間 淳一、 村田 真樹 and 鳥澤 健太郎 (2010). Web データに対応できる日本語異表記対の認定基準. In **言語処理学会 16 回年次大会発表論文集**、 pp. 990–993.
- 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 (2010). 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 **言語処理学会第第 16 回年次大会発表論文集**、 pp. 928–931.
- ALAGIN Forum、 （A-7）日本語異表記対データベース
- ALAGIN Forum、 （A-9）基本的意味関係の事例ベース

Appendix

A.1. 分類器のために使用された素性のリスト

素性 ID	その説明
S1	1 つ目の表記の編集箇所
S2	2 つ目の表記の編集箇所
S3	編集箇所の前方 1 文字
S4	編集箇所の後方 1 文字
S5	編集箇所の前方 2 文字連続
S6	編集箇所の前方 3 文字連続
S7	編集箇所の前方 2 文字目の文字
S8	編集箇所の前方 3 文字目の文字
S9	編集箇所の後方 2 文字連続
S10	編集箇所の後方 3 文字連続
S11	編集箇所の後方 2 文字目の文字
S12	編集箇所の後方 3 文字目の文字
S13	S1-S2 とした文字列
S14	S3-S13 とした文字列
S15	S5-S13 とした文字列
S16	S6-S13 とした文字列
S17	S13-S4 とした文字列
S18	S3-S13-S4 とした文字列
S19	S5-S13-S4 とした文字列
S20	S6-S13-S4 とした文字列
S21	S13-S9 とした文字列
S22	S3-S13-S9 とした文字列
S23	S5-S13-S9 とした文字列
S24	S6-S13-S9 とした文字列
S25	S13-S10 とした文字列
S26	S3-S13-S10 とした文字列
S27	S5-S13-S10 とした文字列
S28	S6-S13-S10 とした文字列
S29	S1、S2、S3、S4 の字種 （漢字、仮名、数字、ローマ字など）
S30	S13、S14、S15、S16、S17、S18 の字種 （漢字、仮名、数字、ローマ字など）

表 4: 使用された素性のリスト 1

素性 ID	その説明
S31	S1、S2、S3、S4 の品詞
S32	S13、S14、S15、S16、S17、S18 の品詞
S33	S1、S2、S3、S4 の品詞と位置情報
S34	S13、S14、S15、S16、S17、S18 の品詞と位置情報
S35	編集箇所の両方が数字の場合、それらが同じ値か否か
S36	編集箇所の両方がひらがなの場合、 それらが大文字と小文字だけの違いか否か
S37	編集箇所の両方がカタカナの場合、 それらが大文字と小文字だけの違いか否か
S38	編集箇所の両方がローマ字の場合、 それらが大文字と小文字だけの違いか否か
S39	編集箇所の両方の字種が一致するか否か
S40	一方の編集箇所に濁点をつけるともう一方の編集箇所になるのか
S41	一方の編集箇所に半濁点をつけるともう一方の編集箇所になるのか
S42	編集箇所が表記対の一方にしかなく、 それが [化、系、類、型、形、氏、ー、・] のどれかの文字なのか
S43	編集箇所が表記対の一方にしかなく、 その表記対の最後の文字と一致するのか
S44	編集箇所が表記対の一方にしかなく、 桁数をあらわす文字なのか (例. 千、万)
S45	編集箇所の字種が漢字とひらがなの場合、 JUMAN 辞書の読みが一致するか否か
S46	編集箇所の両方の字種が漢字の場合、 JUMAN 辞書の読みが一致するか否か

表 5: 使用された素性のリスト 2