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

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† 鳥澤 健太郎‡
† 龍谷大学 t060570@mail.ryukoku.ac.jp, watanabe@rins.ryukoku.ac.jp
‡ 情報通信研究機構 {murata, kazama, kuroda, m-tsuchida, torisawa}@nict.go.jp
§ 公立はこだて未来大学 fujita@fun.ac.jp ¶ 東京大学 eiji.aramaki@gmail.com

1 はじめに

本研究では、検索システムを利用する際に情報の取りこぼしを防いだり、より多くの情報を得たりするために、日本語における編集距離の小さい表記対を対象に、異表記対の抽出を行う。英語表記による略語 [1] やカタカナと英語ペア [2] などの獲得、特定の分野に絞った同義語や異表記の抽出 [3, 4] といった研究はあるが、日本語における表記を対象に、分野を絞らずに機械学習を用いて、異表記対の抽出を行う研究はあまりない。

異表記対の抽出を行うことで、インターネットによる情報検索をする場合、例えば、’スパゲッティ’ の情報を検索したい場合、異表記対を抽出していなければ、’スパゲッティ’ に正確に一致する検索結果のみとなる。’スパゲッティ’ における編集距離の小さい異表記対としては、’スパゲッティー’、’スパゲティ’ がある。異表記対の抽出を行っていなければ、’スパゲッティー’ や’スパゲティ’ と記述されている場合の情報が得られず、情報の取りこぼしが生じてしまう。また、’カワイイ’ の場合、’カワイイ’ における編集距離の小さい異表記対である’カワイィ’ を検索対象に含めることで、’カワイイ’ のみでは得られなかった情報を得ることができる。

実験データを作成する際に、異表記対と判定された表記対の多くが編集距離が 1 であった。異表記対は編集距離が 1 であることが多いと考え、本研究は日本語表記である編集距離の小さい表記対を対象に、機械学習と種々の素性を用いた異表記対の抽出を行った。

2 問題設定と提案手法

2.1 異表記対の定義

本研究において、異表記対の定義について述べる。例えば、(問い合わせメール, 問合わせメール) と (学園闘争, 学園紛争) は編集距離の小さい表記対である。(問い合わせメール, 問合わせメール) は、異表記対になり、(学園闘争, 学園紛争) は、同義語対ではあるが、闘争と紛争が異なる形態素のため異表記対にはならない。表記対が、同義語対であり、同一形態素なら異表記対とし、同一形態素でなければ異表記対でないとする。(詳細は、黒田ら [5] を参照。)

2.2 異表記対の抽出における難しさ

異表記対は、web 上において類似した文脈で出現する。しかし、文脈類似のみの条件に基づく抽出では同義語対や類義語対も抽出してしまう。そこで、日本語表記である編集距離が 1 の表記対に絞って、文脈の類似する表記対を取り出すことで、高い F 値で異表記対を抽出できるのではないかと考えた。風間らが文脈類似度を用いて作成した大規模類似語リスト [6] (100 万表記に対して類似語が最大 500 個与えられているもの) より、編集距離が 1 である類似語対をランダムに取り出し、それらを異表記対であるとし、人手による評価を行った。取り出した異表記対の F 値を求めたところ、0.007 であった。F 値 0.007 は非常に低い値であり、文脈類似と編集距離を用いただけでは、異表記対の抽出を行うことは、難しいことを示している。

2.3 提案手法

そこでわれわれは、前節の問題を解決し、高い F 値で異表記対の抽出を行うために、種々の素性を用いた教師あり機械学習法を用いた。機械学習法として SVM 法を用いた。SVM には TinySVM の線形カーネルを利用し、ソフトマージンパラメータを 1 とした。

素性は、一般的に用いられる’編集箇所の文字’ や、’その周辺の文字’ といった情報を用いた。それだけでは、F 値があまり良好ではなかったため、表記対が日本語の表記であるということに着目した素性を考案した。既存の日本語辞書の情報を用いた素性や、編集箇所の文字や表記対から読み取れる情報を素性とした。素性の詳細は 2.5 節で述べる。

2.4 実験に用いるデータ

実験に用いるデータは、風間らが作成した大規模類似語リスト [6] を基に作成した。大規模類似語リストに含まれる 100 万表記と、その表記の各々にある最大 500 個の類似語を表記対とした。作成した表記対から編集距離が 1 の表記対をランダムに 14,185 組取り出し、異表記対であるのか否かのタグ付けを行った。提案手法に用いる素性には、表記対の各表記が対の 1 つ目にあるか、2 つ目にあるかの違いで、異なる情報になる素性がある。その素性の情報量を増やすために、タグ付けを行った表記対から、1 つ目と 2 つ目の表記を入れ替えた表記対を作成し、合計 28,370 組を実験データとして用いる。なお、順番を入れ替える前の表記対と入れ替えた後の表記対は、同じタグを付与した。

実験データを、データ A とデータ B の 2 つにわけた。異表記対とタグ付けがされた表記対は、データ A には 745 組、データ B には 725 組であった。データ A は素性の考案に利用した。データ B は考案した素性が他のデータでも有効であるかを確認する実験に用いた。素性の考案は、データ A での 10 分割クロスバリデーションの実験結果を考察することにより行った。素性が定まってから、データ A を学習データ、データ B をテストデータとする実験を行った。さらに、提案手法の有効性を確認するために比較実験を行い、素性の有効性をブートストラップ法により確認した。

2.5 提案手法に用いる素性

提案手法に用いる素性を表 1 に示す。S1-S52 は、編集箇所とその周辺の文字の情報を素性にした。なお位置情報とは、対象の文字が形態素のどの位置にあるかの情報である。S53 は、大規模類似語リストを作成した際に、用いられた類似度を素性にした。類似度の値をそのまま使うのではなく、離散的に情報を付与することにした。S54 は、スタッキングアルゴリズムを使用し、JUMAN 辞書において未定義である表記対に対しても、近似的に S64 の情報を付与する。スタッキングアルゴリズムでは、大規模類似語リストから作成した表記対において、それぞれの対となる各表記の JUMAN 辞書に定義されている代表表記が、一致するか否かによりタグ付けを行ったデータを用いる。このデータは、JUMAN 辞書に未定義語を含む表記対は取り除き、データ A、データ B と同じく、対の 1 つ目と 2 つ目の表記を入れ替え、合計 904,612 組のデータとし、S54 以外の表 1 の素性を付与し、学習データとした。この学習データには、JUMAN 辞書の代表表記が一致するとタグ付けされたデータは 25,934 組あった。この学習データと、S54 以外の素性が付与している S54 の素性を付与したいデータをテストデータとする実験を行い、分類結果を S54 の素性を付与したいデータにおける S54 の情報として付与する。JUMAN 辞書に定義されていない表記対においても分類を行うため、S64 の情報を付与できない表記対にも、JUMAN

表 1: 素性

S1	1 つ目の表記の編集箇所	S49	S13 の品詞と位置情報
S2	2 つ目の表記の編集箇所	S50	S14 の品詞と位置情報
S3	編集箇所の前方 1 文字	S51	S17 の品詞と位置情報
S4	編集箇所の後方 1 文字	S52	S18 の品詞と位置情報
S5	編集箇所の前方 2 文字連続	S53	表記対の類似度
S6	編集箇所の前方 3 文字連続	S54	スタッキングアルゴリズムにより、表記対を JUMAN 辞書の代表表記が一致と分類するか否か
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	S55	編集箇所の両方が数字の場合、それらが同じ値か否か
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	S56	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
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	S57	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
S13	S1-S2 とした文字列		
S14	S3-S13 とした文字列	S58	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
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	S59	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
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	S60	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
S21	S13-S7 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	S61	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
S24	S6-S13-S9 とした文字列		
S25	S13-S8 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	S62	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
S28	S6-S13-S10 とした文字列		
S29	S1 の字種		
S30	S2 の字種	S63	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)
S31	S3 の字種		
S32	S4 の字種		
S33	S13 の字種	S64	表記対のそれぞれの表記の JUMAN 辞書の代表表記が、一致するか否か
S34	S14 の字種		
S35	S17 の字種		
S36	S18 の字種	S65	表記対が日本語ワードネット辞書に類義語対として定義されているか否か
S37	S1 の品詞		
S38	S2 の品詞		
S39	S3 の品詞	S66	表記対の編集箇所の文字の組が、異体字辞書に異体字の組として定義されているか否か
S40	S4 の品詞		
S41	S13 の品詞		
S42	S14 の品詞		
S43	S17 の品詞	S67	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
S44	S18 の品詞		
S45	S1 の品詞と位置情報		
S46	S2 の品詞と位置情報	S68	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
S47	S3 の品詞と位置情報		
S48	S4 の品詞と位置情報		

代表表記の情報を付与できる。S55-S63 は、編集距離の小さい表記対の編集箇所に注目した素性である。S55-S60 は、編集箇所が (A,a)、(1, 一)、(や,ゃ)、(か, が) などの類似した文字による置換によって等しい文字列になり、S61-S63 は、編集箇所が削除によって等しい文字列になる表記対が対象である。S64-S66 は、種々の辞書を用いた素性である。種々の辞書で、同じ意味となる表記を種々の辞書における表記対と定義し、それらの表記対と一致するか否かの素性である。S67、S68 は、JUMAN を用いて、それぞれの表記の読みが一致するか否かの素性である。

3 評価実験

3.1 一般的な手法における比較実験

表 1 の素性すべてを用いて、データ A を学習データ、データ B をテストデータとする実験において、提案手法による実験と、すべて異表記対であると判定した場合のベースライン手法 (2.2 節の冒頭の箇所で試した手法と等価である。) と、異表記対を抽出するために考案した素性 (S53-68) を省いた、一般的な素性 (S1-S52) のみを付与した比較手法 A における、正解率と再現率、適合率、F 値を求めた結果を表 2 に示す。ベースラインにおいて、F 値が 0.007 と低い値であり、異表記対における抽出の難しさを示している。比較手法 A において、提案手法の方がすべての値に対して高い値を示しており、S53-S68 といった種々の素性を用いる提案手法の有効性を示している。

表 2: 種々の素性を用いた提案手法による実験結果

手法	正解率	再現率	適合率	F 値
提案手法	99.12%	99.07%	92.29%	0.912
ベースライン手法	5.25%	100%	5.25%	0.007
比較手法 A	98.53%	81.24%	88.97%	0.849
比較手法 B	97.04%	57.52%	78.83%	0.665
比較手法 C	45.31%	23.72%	2.33%	0.042
比較手法 D/E	96.36%	29.52%	97.27%	0.452

表 3: 種々の辞書において提案手法が異表記対に分類した割合

辞書	異表記対に分類された割合	
EDR 辞書	45.15%	(10,800/23,920)
日本語ワードネット	21.36%	(16,814/78,717)
JUMAN 辞書	83.34%	(19,249/23,097)

表 4: 提案手法によるデータ B と種々の辞書における分類結果

	異表記対に分類	非異表記対に分類
データ B	(書いてた頃, 書いていた頃) (自サーバー, 自サーバ) (日光彫, 日光彫り)	(仕上塗材, 上塗材) (Ｌｖ 60, Ｌｖ 60 台) (160 ＧＢ,60 ＧＢ)
EDR 辞書	(竜山文化, 龍山文化) (言い替えする, 言替えする) (爽かだ, 爽快だ)	(前, 前方) (極道だ, 獄道だ) (科目, 目)
日本語ワードネット	(寄せ合せ, 寄せ合わせ) (追尋, 追蹤) (跳ねる, 跳返る)	(異母兄弟, 異父兄弟) (若い者, 若い衆) (凶漢, 悪漢)
JUMAN 辞書	(すじ合, すじ合い) (よきょう, 余きょう) (なげすてる, 投げすてる)	(洞々たる, 洞洞たる) (やけ残る, 焼け残る) (糯米, 餅米)

表 5: 荒牧らの手法と類似した素性

A1	1 つ目の表記の編集箇所→ 2 つ目の表記の編集箇所
A2	編集箇所の前方 1 文字
A3	編集箇所の後方 1 文字
A4	A1 の TYPE(ひらがな、カタカナ、漢字、それ以外)
A5	A2 の TYPE(ひらがな、カタカナ、漢字、それ以外)
A6	A3 の TYPE(ひらがな、カタカナ、漢字、それ以外)
A7	1−(編集距離 ×2÷1 つ目の表記の文字数 +2 つ目の表記の文字数) で求められる値

3.2 種々の辞書との比較

提案手法が編集距離が 1 の異表記対をどのくらい抽出できるのかを、種々の辞書を用いて比較を行った。種々の辞書としては、EDR 辞書、日本語ワードネット辞書、JUMAN 辞書を用いる。編集距離が 1 の表記対は、EDR 辞書には 933,037 組、日本語ワードネット辞書には 78,717 組、JUMAN 辞書には 23,097 組あることがわかった。EDR 辞書には人名が含まれている。本報告では人名は異表記対の対象でないと判断し、取り除いた。その結果、EDR 辞書に含まれる編集距離が 1 の表記対は 23,920 組であった。JUMAN 辞書は同じ代表表記をもつ表記を表記対として扱った。

データ A を学習データとし、データ B と種々の辞書それぞれをテストデータとする提案手法による実験を行った。種々の辞書において、異表記対であると分類した割合を表 3 に示す。データ B と種々の辞書それぞれにおいて、異表記対であると分類された表記対と、分類されなかった表記対をそれぞれランダムに、3 組ずつ取り出した結果を表 4 に示す。

表 4 より、提案手法による種々の辞書における分類は、良好であると判断できる。表 3 において、EDR 辞書と日本語ワードネットの正解率があまり高くないのは、同義語対や類義語対など、異表記対以外のものが多く含まれていたためと思われる。

3.3 先行研究との比較実験

提案手法の効果を確認するために、日本語の表記であり表記対が異表記対か否かを判定する荒牧ら [4] の手法と比較実験を行った。まず、データ A を学習データ、データ B をテストデータとし、荒牧らの手法と表 5 に示す類似した素性を用い、SVM 法による実験を行った。これを比較手法 B とする。次に、荒牧らが公開しているツール[1] を用いて、データ B をテストデータとした実験を行った。これを比較手法 C とする。比較手法 B は、

[1] http://202.218.239.69/~aramaki/TRANS/

本研究で考案した素性の有効性を確認するために行った。比較手法 C は、先行研究に対する優位性を確認するために行った。

比較手法 B と比較手法 C それぞれの正解率と再現率、適合率、F 値を求めた結果を表 2 に示す。比較手法 B、C ともに、提案手法の方が良好な結果となった。比較手法 C では、特に編集箇所が数字の場合に誤った分類をしてしまうことが多く、編集箇所が数字を対象にした学習データを用いていなかったのではないかと考えられる。提案手法が誤り、先行手法が正しい分類をした表記対は、(魚沼産コシヒカリ, 魚沼コシヒカリ) や (米国株市場, 米国株式市場) といった削除による編集が行われるデータが多かった。

3.4 提案手法に用いた素性の検討

提案手法に用いた素性における有意差の分析を行った。すべての素性を用いた場合の出力と、すべての素性から 1 種類の素性を省いた場合の出力を比較し、ブートストラップ法 (反復数 10,000) を用いてそれらの有意差を調べた。データ A のみを利用した 10 分割クロスバリデーションの検討実験 α と、データ A を学習データ、データ B をテストデータとした検討実験 β を行った。

表 6 に全素性 (S1-S68) を用いた場合の F 値が高かった回数 (勝利回数) と、素性 (S53-S68) を 1 種類ずつ省いた場合の勝利回数と、検討実験 β において、全素性の F 値から 1 種類省いた場合の F 値差を示す。検討実験 α、β ともに、すべての素性を用いた場合の勝利回数が 9,500 回を超える素性は、S55、S58、S67 であった。これらの素性が特に有効であり、用いることで F 値を 0.01 以上向上させることがわかった。これらの素性の特徴は、素性の情報がそのまま機械学習の分類結果になる場合が多く、機械学習を用いない方法にも有効である。そこで、S55 が' 同じ値' または、S58 が' 一致' または、S67 が' 一致' のどれかの情報が付与されていれば、異表記対であると判定し、それ以外の場合であれば異表記対でないと判定するという比較手法 D を考案した。さらに、データ A を学習データ、データ B をテストデータとし、SVM 法において S55、S58、S67 の素性のみを用いた比較手法 E の実験を行った。

表 2 に、比較手法 D と比較手法 E における正解率と再現率、適合率、F 値を求めた結果を示す。比較手法 D と E は同じ値であった。同じ値であるのは、学習データにおいて、それぞれの素性が同じ表記対に付与されることがなく、S55 が' 同じ値' または、S58 が' 一致' または、S67 が' 一致' のどれかの情報が付与されていれば、異表記対のタグが付いている場合が多く、それ以外の場合であれば非異表記対のタグが付いている場合が多かったため、比較手法 D と同じ結果となったのではないかと考えられる。提案手法の F 値が比較手法 D、E の F 値よりも高い値であった。勝利回数が 9,500 回を超えない素性であっても、種々の素性を用いている提案手法が有効であることを示している。なお、表 2 において、ブートストラップ法 (反復数 10,000) を用いて、提案手法とベースライン手法や比較手法の F 値における有意差を調べたところ、提案手法の勝利回数が 10,000 回であり、提案手法の有効性を確認した。

S54 と S64 は、ともに JUMAN 辞書の代表表記を用いた類似した素性である。しかし、全素性を用いた検討実験 α、β ともに、S54 を省いた方が S64 を省くよりもよりも、はるかに高い勝利回数となっている。スタッキングアルゴリズムを用い、S64 の情報が付与できなかった表記対にも情報を付与することで、高い勝利回数になったと考えられる。S64 のように情報を付与できないデータがある場合に、スタッキングアルゴリズムを用いることで、性能が高くなることを確認した。

検討実験 α、β ともに、素性を 1 種類省いた場合の勝利回数が 9,500 回を超える素性はなかった。このことから、全素性を用いるのがよいことがわかった。なお、S1-S52 の素性においても、検討実験 α、β ともに、全素性または、素性を 1 種類省いた場合において勝利回数が 9,500 回を超える素性はなかった。

表 6: ブートストラップ法による素性の分析

省いた素性	全素性		1 種類		F 値差
	検討 α	検討 β	検討 α	検討 β	検討 β
S53	9,361	6,206	630	3,424	0.001
S54	9,462	9,682	392	275	0.003
S55	9,996	9,942	4	57	0.011
S56	6,061	0	1,828	0	0
S57	0	0	9,533	0	0
S58	10,000	10,000	0	0	0.015
S59	0	6,314	0	0	0.001
S60	0	0	0	0	0
S61	5,062	215	4,629	9,509	-0.003
S62	6,086	6,302	1,841	0	0.001
S63	8,658	0	0	6,353	0
S64	3,463	0	3,403	0	0
S65	0	0	0	0	0
S66	0	6,341	0	0	0
S67	9,999	10,000	1	0	0.052
S68	8,075	1,024	1,859	8,973	-0.002

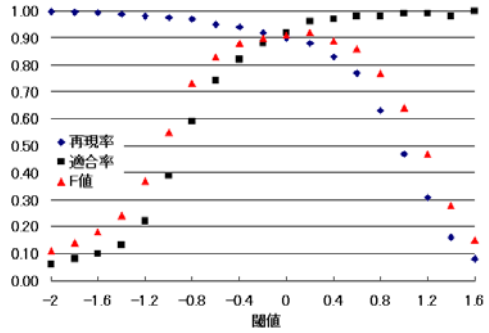

図 1: 閾値を変化させた場合の再現率、適合率、F 値

4 異表記対の抽出

データ A を学習データとし、2.4 節で作成した表記対からデータ A、データ B 以外の、編集距離が 1 の表記対である 1,068,461 組をテストデータとする実験を行った。この実験結果より、異表記対の抽出を行った。

4.1 抽出量の設定

データ A を学習データ、データ B をテストデータとする提案手法による実験を行い、データの分類を識別関数の出力値が正か負で行うのではなく、閾値を設定することで異表記対の抽出量を決定する実験を試みた。

図 1 に閾値を変化させた場合の再現率、適合率、F 値の値を示す。閾値を 0.1 に設定することで、F 値 0.932 と最も高い値が得られた。閾値をさらに高く設定することで、F 値は低くなるが、網羅的に異表記対の抽出を行うことができる。閾値を低く設定することで、抽出できる異表記対は少なくなるが、誤りが少なくすることができる。抽出する目的にあわせて、閾値を設定することができる。

4.2 種々の辞書との重なり

閾値を 0 に設定し、異表記対の抽出を行い、新たに種々の辞書にない異表記対の抽出ができるか調査を行った。

異表記対の抽出を行った結果、159,896 組が抽出された。抽出した異表記対が種々の辞書に含まれている割合と、種々の辞書に含まれていない割合を表 7 に示し、抽出した異表記対にあり種々の辞書にない表記対と、抽出した異表記対になく種々の辞書にある表記対をランダムに、3 対ずつ取り出した結果を表 8 に示す。抽出した異表記対は、ほとんどが既存の辞書になく、提案手法により新たに異表記対を獲得できることがわかった。

5 パターンによる異表記候補対の生成

前節までに、編集距離が 1 の表記対から、異表記対の抽出方法を説明した。3.2 節で、大規模類似語リストから取り出した表記対を、提案手法で異表記対を抽出しただけでは、異表記対をすべて網羅しきれていないことを示した。異表記対を可能な限り網羅するためには、異表記候補対である表記対を増やせばよ

表 7: 抽出した異表記対と種々の辞書の割合

辞書	含まれている割合	含まれていない割合
EDR 辞書	20.68% (4,948/23,920)	96.91% (154,948/159,896)
日本語ワードネット	1.79% (1,414/78,717)	99.12% (158,482/159,896)
JUMAN 辞書	6.59% (1,523/23,097)	99.05% (158,373/159,896)

表 8: 抽出した異表記対と種々の辞書にのみある表記対

辞書	異表記対にのみあるデータ	種々の辞書にのみあるデータ
EDR 辞書	(会計帳簿等, 会計帳簿)	(お冠, 御冠)
	(行ったせい, いったせい)	(縄付, 縄付き)
	(ユリアン, リアン)	(雁, 鴻雁)
日本語ワードネット	(凹凸部, 凹凸部分)	(継接, 継接ぎ)
	(期末試験, 学期末試験)	(締切り, 締切日)
	(身の上話, 身の上)	(禁制, 禁断)
JUMAN 辞書	(ログ内容, ブログ内容)	(飛び越す, とび越す)
	(ディルドー, ディルドウ)	(吹き飛ばす, 吹飛ばす)
	(買物帰り, 買い物帰り)	(釣さげる, 釣りさげる)

表 9: 異表記候補対の生成に用いたパターン

	編集文字	例	頻度
上位	[・,del]	(塩コショウ, 塩・コショウ)	102
	[ー,del]	(有線ルータ, 有線ルーター)	88
	[い,del]	(乗っていた時, 乗ってた時)	78
下位	[六,6]	(六メートル,6 メートル)	1
	[篭, こ]	(引き篭もり, 引きこもり)	1
	[っ,del]	(引っ張り応力, 引張り応力)	1

いと考えた。そこで、異表記対になりやすいパターンを用いて、表記から異表記候補対の生成を行った。

異表記対になりやすいパターンは、データ A、データ B から 217 個取り出した。表 9 に取り出したパターンの出現頻度上位 3 個と下位 3 個を示す。これらは、異表記対であるとタグ付けされた 1,470 組の表記対の編集箇所に着目して、取り出した。del は、もう一方の編集文字を削除することを示す。検索エンジン研究基盤 TSUBAKI [7] に出現する係り受けが 2 種類以上ある名詞が約 3,000 万表記あった。その約 3,000 万表記に取り出したパターンを適用させて異表記対を生成する。パターンを適用させる方法としては、約 3,000 万表記に、取り出したパターンに対応する編集文字があるかを調べ、あればその表記を表記 A とし、表記 A をパターン (例. 表記 A:リモート・マシン→表記 B:リモートマシン、リモト・マシン) に適用し、生成した表記 B と組み合わせ、表記対とする。また、編集文字が del を含むパターンにおいては、表記 A が、表記 B からもう一方の編集文字を削除した表記の場合がある。この場合を考え、表記の文字と文字の間に、del のもう一方の編集文字を挿入し、表記対とする。この方法を約 3,000 万表記すべてに適用した場合、膨大な数の表記対が生成されてしまうと考え、パターンに適用した表記 B が、約 3,000 万表記に含まれていない表記対は取り除いた。その結果、8,311,219 組の異表記候補対が生成された。

異表記候補対がどれくらい網羅率があるのかを確認するために、種々の辞書から約 3,000 万表記に含まれない表記対を取り除き、網羅率を調べた結果を表 10 に示す。種々の辞書における網羅率はあまり高くなく、分子の個数 (EDR 辞書、JUMAN 辞書) が表 7 における抽出した異表記対が種々の辞書に含まれている個数より減っていた。これは、異表記対であるタグ付けされた表記対から取り出した 217 組のパターンが、少なかったためと考えらる。新たなパターンの候補としては、種々の辞書における表記対の編集箇所に着目して取り出そうと考えている。

データ A を学習データ、異表記候補対をテストデータとし、提案手法により異表記対の抽出を行ったところ、4,220,686 組が抽出できた。抽出した表記対をランダムに 30 組取り出し、人手でチェックをした結果、21 組 (70%) が正しく抽出されていた。おおよそではあるが、異表記候補対から 280 万組の異表記対が、正しく抽出できたのではないかと考えられる。異表記候補対から、種々の辞書になく、異表記対であると人手で判断した表記対の一部を表 11 に示す。

表 10: 異表記候補対の種々の辞書に対する網羅率

辞書	生成した表記対の網羅率	
EDR 辞書	26.17%	(2,662/10,177)
日本語ワードネット	9.62%	(5,023/52,209)
JUMAN 辞書	27.85%	(1,415/5,080)

表 11: 異表記候補対から提案手法により抽出した異表記対

(ホーエンツォレルン家, ホーエンツオレルン家)	(リクィッドタイプ, リクイッドタイプ)
(茶の間風, お茶の間風)	(ハルマゲドン, ハーマゲドン)
(思ったくらい, 思ったぐらい)	(イノウ, イノー)
(千五百円, 千 5 百円)	(フェーブ, フェーヴ)
(ラムズフェルド国防長官, ラムスフェルド国防長官)	(カリフォルニア州法, カルフォルニア州法)
(ポートランド, ポルトランド)	(ヒラドツツジ, ヒラトツツジ)
(マグロ振り, マグロぶり)	(PukiWIkiMod,PukiWikiMod)
(トランスパックレース, トランパックレース)	(キュレイション, キュレーション)
(青山 1 丁目駅, 青山一丁目駅)	(リーダースチョイス, リーダーズチョイス)

6 おわりに

本研究は、機械学習と種々の素性を用いて、日本語の表記を対象に編集距離が小さい異表記対の抽出を行った。web 上から文脈類似と編集距離が 1 の条件に基づく抽出、種々の辞書や先行研究と比較を行ったところ、提案手法の効果を確認した。

提案手法に用いた種々の素性の検討を行ったところ、特に有効な素性があり、それらは F 値を 0.01 以上向上させた。さらに、素性の情報が付与できないデータに対して、スタッキングアルゴリズムを使用することで、近似的ではあるが、情報を付与することができ、スタッキングアルゴリズムを使用しない素性よりもわずかではあるが、F 値を向上させることを確認した。

考案した素性 (S53-S68) が有効か確認するために、一般的な素性 (S1-S52) のみを用いた手法による実験したところ、考案した素性を用いることで F 値を 0.063(0.849 → 0.912) 向上することを確認した。

提案手法により、大規模類似語リストから異表記対を抽出した 95%以上が、種々の辞書にない新たな異表記対であった。パターンを用いた超大規模類似語リストの生成を行った。434 組のパターンから約 830 万組の表記対が生成できた。提案手法で超大規模類似語リストから異表記対を抽出したところ、おおよそではあるが、280 万組の異表記対を正しく抽出した。提案手法で抽出した異表記対は 95%以上が、種々の辞書にない新たな異表記対であったことより、超大規模類似語リストから、おおよそ 250 万組が正しく抽出でき、種々の辞書にない異表記対ではないかと考えられる。

参考文献

[1] 岡崎直観, 辻井潤一: “アライメント識別モデルを用いた略語定義の自動獲得”, 言語処理学会第 14 回年次大会発表論文集, pp.139-142, (2008).

[2] Eric Brill: “Automatically harvesting Katakana-English term pairs from search engine query logs”, Proc. Sixth Natural Language Processing Pacific Rim Symposium, pp.393-399 ,(2001).

[3] Yoshimasa Tshruoka, et al.: “Learning string similarity measures for gene/protein name dictionary look-up using logistic regression”, Bioinformatics, 23(20):2768-74, (2007).

[4] Eiji Aramaki, Takeshi Imai, Kengo Miyo and Kazuhiko Ohe: “Orthographic Disambiguation Incorporating Transliterated Probability”, IJCNLP2008, (2008).

[5] 黒田航, 風間淳一, 村田真樹, 鳥澤健太郎: “Web 文書にも対応できる日本語異表記の認定基準”, 言語処理学会第 16 回年次大会発表論文集, (2010).

[6] 風間淳一, De Saeger, Stijn, 鳥澤健太郎, 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”, 言語処理学会第 15 回年次大会発表論文集, pp.84-87, (2009).

[7] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. Thubaki: “An open search engine infrastructure for developing new information access.”, IJCNLP 2008, (2008).