(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2011-138440
(P2011-138440A)
(43)公開日　平成23年7月14日(2011.7.14)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/27	(2006.01)	G06F	17/27	Z	5B075
G06F 17/30	(2006.01)	G06F	17/30	320D	5B091
		G06F	17/30	170A	

審査請求 未請求　請求項の数 13　ＯＬ　（全 48 頁）

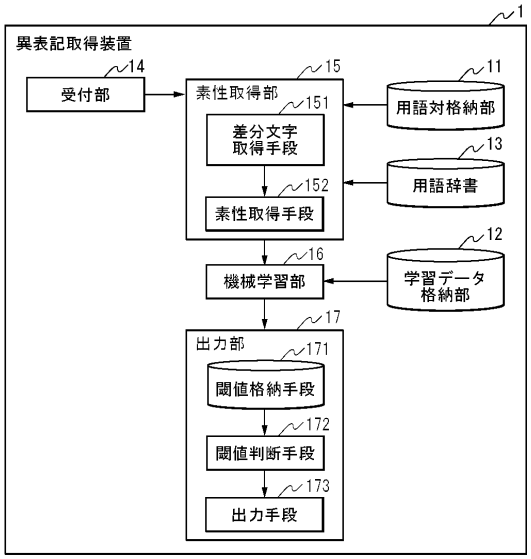
(21)出願番号　特願2009-299287 (P2009-299287)
(22)出願日　平成21年12月30日 (2009.12.30)

(71)出願人　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100115749
弁理士　谷川　英和
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　小島　正裕
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　鳥澤　健太郎
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

最終頁に続く

(54)【発明の名称】異表記取得装置、異表記取得方法、およびプログラム

(57)【要約】
【課題】従来、十分な異表記抽出の精度が得られなかった。
【解決手段】用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納し、編集距離が１以上の用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する素性取得部と、用語対に対して、２以上の学習データと取得された複数の素性とを用いて、教師あり機械学習法により、各用語対が異表記の用語対であるか否かを判断する機械学習部と、判断結果を出力する出力部とを具備する異表記取得装置により、精度の高い異表記の用語対抽出ができる。
【選択図】図１



【特許請求の範囲】
【請求項１】
編集距離が１以上の用語対を１以上格納し得る用語対格納部と、
用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、前記用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、前記用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納し得る学習データ格納部と、
前記用語対格納部の用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する素性取得部と、
前記各用語対に対して、前記学習データ格納部の２以上の学習データと、前記素性取得部が取得した複数の素性とを用いて、教師あり機械学習法により、前記用語対格納部の各用語対が異表記の用語対であるか否かを判断する機械学習部と、
前記機械学習部における判断結果を出力する出力部とを具備する異表記取得装置。
【請求項２】
前記字種関連素性は、
用語対が有する２つの用語の編集箇所の字種が異なり、かつ、当該２つの用語の編集箇所が数字であるか否かを示す情報であり、
前記素性取得部は、
前記用語対格納部の用語対ごとに、用語対が有する２つの用語の編集箇所の字種が異なり、かつ、当該２つの用語の編集箇所が同じ値の数字であるという条件に合致するか否かを判断し、当該判断結果を字種関連素性として取得する請求項１記載の異表記取得装置。
【請求項３】
前記字種関連素性は、
用語対が有する２つの用語の編集箇所の字種がローマ字であり、かつ、当該２つの用語の編集箇所が大文字と小文字の違いであるか否かを示す情報であり、
前記素性取得部は、
前記用語対格納部の用語対ごとに、用語対が有する２つの用語の編集箇所の字種がローマ字であり、かつ、当該２つの用語の編集箇所が大文字と小文字の違いであるという条件に合致するか否かを判断し、当該判断結果を字種関連素性として取得する請求項１記載の異表記取得装置。
【請求項４】
用語と、当該用語の代表表記とを有する１以上の用語情報を格納し得る用語辞書をさらに具備し、
前記辞書関連素性は、
用語対が有する２つの用語の代表表記が同一であるか否かを示す情報であり、
前記素性取得部は、
前記用語対格納部の用語対ごとに、用語対が有する２つの用語の代表表記を、前記用語辞書から取得し、当該取得した２つの代表表記が同一であるか否かを判断し、当該判断結果を辞書関連素性として取得する請求項１記載の異表記取得装置。
【請求項５】
前記辞書関連素性は、
スタッキングアルゴリズムを使用して、前記教師あり機械学習法とは異なる分類方法、または同一の分類方法であるが学習データが異なる分類方法により、用語対が異表記の用語対であるか否かを判断した結果であり、
前記素性取得部は、
前記用語対格納部の用語対ごとに、前記教師あり機械学習法とは異なる分類方法、または同一の分類方法であるが学習データが異なる分類方法により、当該用語対が異表記の用語対であるか否かを判断し、当該判断結果を辞書関連素性として取得する請求項１記載の異表記取得装置。

【請求項６】
用語と、当該用語の読みとを有する１以上の用語情報を格納し得る用語辞書をさらに具備し、
前記辞書関連素性は、
用語対が有する２つの用語の読みが一致するか否かを示す情報であり、
前記素性取得部は、
前記用語対格納部の用語対ごとに、前記用語辞書から前記用語対が有する２つの用語の読みを取得し、当該２つの用語の読みが一致するか否かを判断し、当該判断結果を辞書関連素性として取得する請求項１記載の異表記取得装置。
【請求項７】
前記機械学習部は、
前記用語対格納部の各用語対が異表記の用語対であるか否かを判断するとともに、異表記の用語対である確度を示すスコアも取得し、
前記出力部は、
前記機械学習部が取得したスコアを出力する請求項１から請求項６いずれか記載の異表記取得装置。
【請求項８】
前記出力部は、
スコアの閾値を格納している閾値格納手段と、
前記機械学習部が取得したスコアが前記閾値以上または前記閾値より大きいか否かを判断する閾値判断手段と、
前記閾値判断手段が前記閾値以上または前記閾値より大きいと判断したスコアに対応する用語対を、異表記の用語対であるとの判断結果とし、当該判断結果または異表記の用語対または異表記でない用語対のいずれか１以上を出力する出力手段とを具備する請求項７記載の異表記取得装置。
【請求項９】
用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、前記用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、前記用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納し得る学習データ格納部と、
異表記のパターンを示す第一文字列と第二文字列とを対に有する異表記パターンを１以上格納し得る異表記パターン格納部と、
１以上の用語を受け付ける受付部と、
前記受付部が受け付けた１以上の各用語に対して、前記異表記パターン格納部の１以上の各異表記パターンを適用し、１以上の用語を生成し、前記１以上の各用語と前記生成した用語とを有する１以上の異表記の候補の用語対である異表記候補用語対を生成する用語対生成部と、
前記用語対生成部が生成した１以上の異表記候補用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上の素性を含む複数の素性を取得する素性取得部と、
前記用語対生成部が生成した各異表記候補用語対に対して、前記学習データ格納部の２以上の学習データと、前記素性取得部が取得した複数の素性とを用いて、教師あり機械学習法により、前記用語対格納部の各異表記候補用語対が異表記の用語対であるか否かを判断する機械学習部と、
前記機械学習部における判断結果を出力する出力部とを具備する異表記取得装置。
【請求項１０】
編集距離が１の異表記の用語対を１以上格納し得る異表記用語対格納部と、
前記異表記用語対格納部に格納されている１以上の異表記の用語対の編集箇所を取得する編集箇所取得部と、
前記編集箇所取得部が取得した編集箇所から、第一文字列と第二文字列とを対に有する異

表記パターンを取得する異表記パターン取得部と、
前記異表記パターン取得部が取得した異表記パターンを、前記異表記パターン格納部に蓄積する異表記パターン蓄積部とをさらに具備する請求項９記載の異表記取得装置。
【請求項１１】
前記用語対の編集距離は２であり、
前記素性取得部は、
前記用語対の２つの差分文字の組を、それぞれ取得する差分文字取得手段と、
前記差分文字取得手段が取得した２つの差分文字を、独立に対象として、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を、２組取得する素性取得手段とを具備し、
前記機械学習部は、
前記素性取得手段が取得した２組の複数の素性のうちの組ごとに、当該各組の複数の素性と、前記学習データ格納部の２以上の学習データとを用いて、教師あり機械学習法により、前記用語対格納部の各組の複数の素性が異表記の用語対に対応する素性の組であるか否かを判断し、当該２つの判断結果を用いて、編集距離が２である用語対が異表記の用語対であるか否かを判断する請求項１から請求項１０いずれか記載の異表記取得装置。
【請求項１２】
記憶媒体に、
編集距離が１以上の用語対、および、
用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、前記用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、前記用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納しており、
素性取得部、機械学習部、および出力部により実現される異表記取得方法であって、
前記素性取得部により、前記記憶媒体の用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する素性取得ステップと、
前記機械学習部により、前記各用語対に対して、前記記憶媒体の２以上の学習データと、前記素性取得ステップで取得された複数の素性とを用いて、教師あり機械学習法により、前記記憶媒体の各用語対が異表記の用語対であるか否かを判断する機械学習ステップと、
前記出力部により、前記機械学習ステップにおける判断結果を出力する出力ステップとを具備する異表記取得方法。
【請求項１３】
記憶媒体に、
編集距離が１以上の用語対、および、
用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、前記用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、前記用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納しており、
コンピュータを、
前記記憶媒体の用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する素性取得部と、
前記各用語対に対して、前記記憶媒体の２以上の学習データと、前記素性取得部が取得した複数の素性とを用いて、教師あり機械学習法により、前記記憶媒体の各用語対が異表記の用語対であるか否かを判断する機械学習部と、
前記機械学習部における判断結果を出力する出力部として機能させるためのプログラム。
【発明の詳細な説明】
【技術分野】
【０００１】

本発明は、異表記の用語対を取り出す異表記取得装置等に関するものである。

【背景技術】

【０００２】

従来の異表記を取り出す技術としては、荒牧らの研究がある（非特許文献１参照）。この研究は、医療分野の専門用語の異表記の取り出しを行うものであった。なお、異表記とは、例えば「スパゲティ」に対して、「スパゲッティ」など、同義語であるが異なる表現の用語を言う。また、異表記の二つの用語を異表記対という。

【０００３】

異表記対の第一の考え方は、以下である。例えば、用語対の例１（問い合わせメール,問合わせメール)、例２(学園闘争,学園紛争)について、例１は異表記対とし、例２は、異表記対ではなく日本語同義語対とする。つまり、第一の考え方において、同一語の異形なら異表記対とし、同一語でなければ、例え意味が同等でも異表記対としない。闘争と紛争は、ほぼ同等の意味を有するが、同一の語でないので、例２は異表記対とはしない。一方、例１の「問い合わせ」「問合わせ」は、表記は異なるが同一の語の異形と判断できるので、異表記対とする。

【０００４】

また、異表記対の第二の考え方は、同義語も異表記とする考え方である。第二の考え方では、上記の例１だけではなく、例２(学園闘争,学園紛争)も異表記対となる。

【０００５】

さらに、異表記、異表記対の考え方は、上記の考え方と類似する考え方でも良く、異表記、異表記対は広く解するものとする。

【０００６】

また、従来技術として、機械学習法についての技術がある(例えば、非特許文献２～非特許文献４参照)

【先行技術文献】

【非特許文献】

【０００７】

【非特許文献１】Eiji Aramaki, Takeshi Imai, Kengo Miyo, Kazuhiko Ohe: Orthographic Disambiguation Incorporating Transliterated Probability, International Joint Conference on Natural Language Processing (IJCNLP2008), pp.48-55, 2008.

【非特許文献２】村田真樹,機械学習に基づく言語処理,龍谷大学理工学部.招待講 演.2004.http://www2.nict.go.jp/x/x161/member/murata/ps/kougi-ml-siryou-new2.pdf

【非特許文献３】サポートベクトルマシンを用いたテンス・アスペクト・モダリティの日英翻訳,村田真樹,馬青,内元清貴,井佐原均,電子情報通信学会言語理解とコミュニケーション研究会 NLC2000-78 ,2001年.

【非特許文献４】SENSEVAL2J辞書タスクでのCRLの取り組み,村田真樹,内山将夫,内元清貴,馬青,井佐原均,電子情報通信学会言語理解とコミュニケーション研究会 NLC2001-40 ,2001年.

【発明の概要】

【発明が解決しようとする課題】

【０００８】

しかしながら、従来技術においては、日本語の一般の異表記を扱うものではなく、また、従来技術を日本語の一般の異表記に適用したのでは、十分な異表記抽出の精度が得られなかった。

【課題を解決するための手段】

【０００９】

本第一の発明の異表記取得装置は、編集距離が１以上の用語対を１以上格納し得る用語対格納部と、用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、用語対が

異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納し得る学習データ格納部と、用語対格納部の用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する素性取得部と、各用語対に対して、学習データ格納部の２以上の学習データと、素性取得部が取得した複数の素性とを用いて、教師あり機械学習法により、用語対格納部の各用語対が異表記の用語対であるか否かを判断する機械学習部と、機械学習部における判断結果を出力する出力部とを具備する異表記取得装置である。

【００１０】

かかる構成により、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる。

【００１１】

また、本第二の発明の異表記取得装置は、第一の発明に対して、字種関連素性は、用語対が有する２つの用語の編集箇所の字種が異なり、かつ、２つの用語の編集箇所が数字であるか否かを示す情報であり、素性取得部は、用語対格納部の用語対ごとに、用語対が有する２つの用語の編集箇所の字種が異なり、かつ、２つの用語の編集箇所が同じ値の数字であるという条件に合致するか否かを判断し、判断結果を字種関連素性として取得する異表記取得装置である。

【００１２】

かかる構成により、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる。

【００１３】

また、本第三の発明の異表記取得装置は、第一の発明に対して、字種関連素性は、用語対が有する２つの用語の字種がローマ字であり、かつ、２つの用語の編集箇所が大文字と小文字の違いであるか否かを示す情報であり、素性取得部は、用語対格納部の用語対ごとに、用語対が有する２つの用語の編集箇所の字種がローマ字であり、かつ、２つの用語の編集箇所が大文字と小文字の違いであるという条件に合致するか否かを判断し、判断結果を字種関連素性として取得する異表記取得装置である。

【００１４】

かかる構成により、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる。

【００１５】

また、本第四の発明の異表記取得装置は、第一の発明に対して、用語と、用語の代表表記とを有する１以上の用語情報を格納し得る用語辞書をさらに具備し、辞書関連素性は、用語対が有する２つの用語の代表表記が同一であるか否かを示す情報であり、素性取得部は、用語対格納部の用語対ごとに、用語対が有する２つの用語の代表表記を、用語辞書から取得し、取得した２つの代表表記が同一であるか否かを判断し、判断結果を辞書関連素性として取得する異表記取得装置である。

【００１６】

かかる構成により、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる。

【００１７】

また、本第五の発明の異表記取得装置は、第一の発明に対して、辞書関連素性は、スタッキングアルゴリズムを使用して、教師あり機械学習法とは異なる分類方法、または同一の分類方法であるが学習データが異なる分類方法により、用語対が異表記の用語対であるか否かを判断した結果であり、素性取得部は、用語対格納部の用語対ごとに、教師あり機械学習法とは異なる分類方法、または同一の分類方法であるが学習データが異なる分類方法により、用語対が異表記の用語対であるか否かを判断し、判断結果を辞書関連素性として取得する異表記取得装置である。

【００１８】

かかる構成により、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能と

なる。

【００１９】

また、本第六の発明の異表記取得装置は、第一の発明に対して、用語と、用語の読みとを有する１以上の用語情報を格納し得る用語辞書をさらに具備し、辞書関連素性は、用語対が有する２つの用語の読みが一致するか否かを示す情報であり、素性取得部は、用語対格納部の用語対ごとに、用語辞書から用語対が有する２つの用語の読みを取得し、２つの用語の読みが一致するか否かを判断し、判断結果を辞書関連素性として取得する異表記取得装置である。

【００２０】

かかる構成により、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる。

【００２１】

また、本第七の発明の異表記取得装置は、第一から第六いずれかの発明に対して、機械学習部は、用語対格納部の各用語対が異表記の用語対であるか否かを判断するとともに、異表記の用語対である確度を示すスコアも取得し、出力部は、機械学習部が取得したスコアを出力する異表記取得装置である。

【００２２】

かかる構成により、用語対の分野を問わず、さらに精度の高い異表記の用語対の抽出が可能となる。

【００２３】

また、本第八の発明の異表記取得装置は、第七の発明に対して、出力部は、スコアの閾値を格納している閾値格納手段と、機械学習部が取得したスコアが閾値以上または閾値より大きいか否かを判断する閾値判断手段と、閾値判断手段が閾値以上または閾値より大きいと判断したスコアに対応する用語対を、異表記の用語対であるとの判断結果とし、判断結果または異表記の用語対または異表記でない用語対のいずれか１以上を出力する出力手段とを具備する異表記取得装置である。

【００２４】

かかる構成により、用語対の分野を問わず、さらに精度の高い異表記の用語対の抽出が可能となる。

【００２５】

また、本第九の発明の異表記取得装置は、用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納し得る学習データ格納部と、異表記のパターンを示す第一文字列と第二文字列とを対に有する異表記パターンを１以上格納し得る異表記パターン格納部と、１以上の用語を受け付ける受付部と、受付部が受け付けた１以上の各用語に対して、異表記パターン格納部の１以上の各異表記パターンを適用し、１以上の用語を生成し、１以上の各用語と生成した用語とを有する１以上の異表記の候補の用語対である異表記候補用語対を生成する用語対生成部と、用語対生成部が生成した１以上の異表記候補用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上の素性を含む複数の素性を取得する素性取得部と、用語対生成部が生成した各異表記候補用語対に対して、学習データ格納部の２以上の学習データと、素性取得部が取得した複数の素性とを用いて、教師あり機械学習法により、用語対格納部の各異表記候補用語対が異表記の用語対であるか否かを判断する機械学習部と、機械学習部における判断結果を出力する出力部とを具備する異表記取得装置である。

【００２６】

かかる構成により、異表記の用語対の候補を自動生成できる。

【００２７】

また、本第十の発明の異表記取得装置は、第九の発明に対して、編集距離が１の異表記

の用語対を１以上格納し得る異表記用語対格納部と、異表記用語対格納部に格納されている１以上の異表記の用語対の編集箇所を取得する編集箇所取得部と、編集箇所取得部が取得した編集箇所から、第一文字列と第二文字列とを対に有する異表記パターンを取得する異表記パターン取得部と、異表記パターン取得部が取得した異表記パターンを、異表記パターン格納部に蓄積する異表記パターン蓄積部とをさらに具備する異表記取得装置である。

【００２８】

かかる構成により、異表記の用語対の候補を自動生成するための異表記パターンを自動的に取得できる。

【００２９】

また、本第十一の発明の異表記取得装置は、第一から第十いずれかの発明に対して、用語対の編集距離は２であり、素性取得部は、用語対の２つの差分文字の組を、それぞれ取得する差分文字取得手段と、差分文字取得手段が取得した２つの差分文字を、独立に対象として、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を、２組取得する素性取得手段とを具備し、機械学習部は、素性取得手段が取得した２組の複数の素性のうちの組ごとに、各組の複数の素性と、学習データ格納部の２以上の学習データとを用いて、教師あり機械学習法により、用語対格納部の各組の複数の素性が異表記の用語対に対応する素性の組であるか否かを判断し、２つの判断結果を用いて、編集距離が２である用語対が異表記の用語対であるか否かを判断する異表記取得装置である。

【００３０】

かかる構成により、編集距離が２の用語対でも、精度高く、異表記の用語対であるか否かを判断できる。

【発明の効果】

【００３１】

本発明による異表記取得装置によれば、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる。

【図面の簡単な説明】

【００３２】

【図１】本発明の実施の形態１における異表記取得装置のブロック図

【図２】同異表記取得装置の動作について説明するフローチャート

【図３】同素性取得処理の動作について説明するフローチャート

【図４】同用語辞書の例を示す図

【図５】同サポートベクトルマシン法のマージン最大化の概念を示す図

【図６】同実験で用いた編集距離が１の日本語用語対の中に、多数決により日本語異表記対であるか日本語異表記対でないかを判定した内訳を示す図

【図７】同Ｌａｎｄｉｓらによる一致度の評価方法を示す図

【図８】同クローズドデータとオープンデータに対して、ベースラインの手法を適用した結果を示す図

【図９】同ブートストラップ法を用いて素性が有効であるかどうかの検討をした結果を示す図

【図１０】同素性の例を示す図

【図１１】同提案手法を用い、大規模類似語リストから編集距離が１の日本語異表記対と分類された用語対が、種々の辞書にどの程度の割合で含まれているかの検討結果を示す図

【図１２】同ＳＶＭの分類精度を示す図

【図１３】同種々の辞書と用語対ＤＢにおいて、編集距離が１の日本語異表記対であると分類された日本語用語対と、分類されなかった日本語用語対をそれぞれランダムに、５組ずつ取り出した結果を示す図

【図１４】同閾値の評価基準を示す図

【図１５】同再現率と適合率の比率を示す図

【図１６】同ベースライン手法を用いた場合の実験結果を示す図

【図１７】同ベースライン手法を用いた場合の実験結果を示す図
【図１８】同ベースライン手法を用いた場合の実験結果を示す図
【図１９】本発明の実施の形態２における異表記取得装置のブロック図
【図２０】同異表記取得装置の動作について説明するフローチャート
【図２１】同異表記パターンの例を示す図
【図２２】上記実施の形態におけるコンピュータシステムの概観図
【図２３】同コンピュータシステムのブロック図
【発明を実施するための形態】
【００３３】
以下、異表記取得装置等の実施形態について図面を参照して説明する。なお、実施の形態において同じ符号を付した構成要素は同様の動作を行うので、再度の説明を省略する場合がある。
【００３４】
（実施の形態１）
【００３５】
本実施の形態において、編集距離が１または２以上の用語対から、異なる文字の字種に関する素性、用語辞書を用いて取得された素性、２つの用語の類似度のうちの１以上の素性を含む複数の素性を取り出し、当該複数の素性を用いて、用語対が異表記の用語対であるか否かを、教師あり機械学習法により判断する異表記取得装置について説明する。
【００３６】
図１は、本実施の形態における異表記取得装置１のブロック図である。
異表記取得装置１は、用語対格納部１１、学習データ格納部１２、用語辞書１３、受付部１４、素性取得部１５、機械学習部１６、出力部１７を備える。素性取得部１５は、差分文字取得手段１５１、素性取得手段１５２を備える。出力部１７は、閾値格納手段１７１、閾値判断手段１７２、出力手段１７３を備える。
【００３７】
用語対格納部１１は、編集距離が１または２以上の用語対を１以上格納し得る。編集距離とは、異なる文字の数である。また、用語対とは、２つの用語である。編集距離が２の用語対は、異なる文字数が２つの用語である。なお、用語とは、通常、名詞や名詞句であるが、形容詞等の他の品詞の用語でも良い。
【００３８】
学習データ格納部１２は、２以上の学習データを格納し得る。学習データは、用語対の複数の素性と正負情報とを有する。学習データは、用語対を有しても良い。用語対の複数の素性は、ここでは、字種関連素性、辞書関連素性、類似度素性のうちの一以上の素性を含む、とする。なお、素性とは、異表記取得装置１が学習する際に手掛かりとする情報のことである。
【００３９】
字種関連素性とは、用語対の異なる文字である編集箇所の字種に関する素性である。字種関連素性は、例えば、用語対が有する２つの用語の編集箇所の字種が異なり、かつ、２つの用語の編集箇所が同じ値の数字であるか否かを示す情報である。また、字種関連素性は、例えば、用語対が有する２つの用語の文字数が同数であり、かつ、２つの用語の編集箇所の字種が異なり、かつ、２つの用語の編集箇所が同じ値の数字であるか否かを示す情報である。また、字種関連素性は、例えば、用語対が有する２つの用語の編集箇所の字種がローマ字であり、かつ、２つの用語の編集箇所が大文字と小文字の違いであるか否かを示す情報である。また、字種関連素性は、例えば、用語対が有する２つの用語の文字数が同数であり、かつ、２つの用語の編集箇所の字種がローマ字であり、かつ、２つの用語の編集箇所が大文字と小文字の違いであるか否かを示す情報である。
【００４０】
また、辞書関連素性とは、用語辞書１３を用いて取得された素性である。辞書関連素性は、例えば、スタッキングアルゴリズムを使用して、機械学習部１６が利用する教師あり

機械学習法とは異なる分類方法、または同一の分類方法であるが学習データが異なる分類方法により、用語対が異表記の用語対であるか否かを判断した結果である。ここで、「学習データが異なる」とは、学習データの元になる用語対の集合が異なる場合、学習データが有する素性が異なる場合などがある。また、辞書関連素性は、例えば、用語対が有する２つの用語の代表表記が同一であるか否かを示す情報である。また、辞書関連素性は、例えば、用語対が有する２つの用語の読みが一致するか否かを示す情報である。また、辞書関連素性は、例えば、用語対が有する２つの用語の文字数が同数であり、かつ、２つの用語の読みが一致するか否かを示す情報である。なお、分類方法とは、異表記の用語対であるか否かの分類の方法である。また、教師あり機械学習法とは異なる分類方法とは、分類のやり方、アルゴリズムが教師あり機械学習法とは異なることである。

【００４１】

また、類似度素性とは、用語対を構成する２つの用語の類似度を示す素性である。二つの用語の類似度は、それらの用語がＷｅｂ上でよく似た文脈に出現するかどうかの情報を利用して求める。なお、用語の類似度を取得する技術は、「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(NLP2009)」等に記載されている。つまり、２つの用語の類似度の取得方法は公知技術である。２つの用語の類似度の算出方法は問わない。

【００４２】

また、正負情報とは、用語対が異表記の用語対であるか否かを示す情報である。正負情報は、異表記の用語対であれば正例（例えば「１」）、異表記の用語対でなければ負例（例えば「０」）である。

【００４３】

また、用語辞書とは、異表記の用語の情報を含む情報群である。用語辞書の例やデータ構造の例については後述する。

【００４４】

また、他の素性として、編集箇所の文字または編集箇所の文字の周辺の文字の情報である編集箇所文字素性がある。

【００４５】

用語辞書１３は、１以上の用語情報を格納し得る。用語辞書１３は、例えば、異表記の用語の情報を含む情報群である。用語辞書１３は、異表記の２つの用語が、陽に対応付けられている必要はない。用語情報は、例えば、用語と用語の代表表記とを有する。用語情報は、例えば、用語と、用語の読みとを有する。

【００４６】

受付部１４は、ユーザからの入力を受け付ける。この入力とは、例えば、異表記取得装置１を動作させるための動作指示である。受付部１４は、異表記であるか否かを判断する対象の用語対を受け付けても良い。動作指示などの入力手段は、テンキーやキーボードやマウスやメニュー画面によるもの等、何でも良い。受付部１４は、テンキーやキーボード等の入力手段のデバイスドライバーや、メニュー画面の制御ソフトウェア等で実現され得る。

【００４７】

素性取得部１５は、用語対格納部１１の用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する。複数の素性とは、例えば、後述する６８の素性である。

【００４８】

素性取得部１５は、用語対格納部１１の用語対ごとに、用語対が有する２つの用語の編集箇所の字種が異なり、かつ、２つの用語の編集箇所が同じ値の数字であるという条件に合致するか否かを判断し、判断結果を字種関連素性として取得する。なお、素性取得部１５は、例えば、用語対「３者会談」「三者会談」に対して、編集箇所が「３」「三」であるので、上記条件に合致する、と判断する。また、素性取得部１５は、例えば、用語対「１２５件」「百二十五件」に対して、編集箇所が「１２５」「百二十五」であるので、用

語の文字数は同数ではないが、上記条件に合致する、と判断する。

【００４９】

素性取得部１５は、例えば、用語対格納部１１の用語対ごとに、用語対が有する２つの用語の編集箇所の字種がローマ字であり、かつ、２つの用語の編集箇所が大文字と小文字の違いであるという条件に合致するか否かを判断し、判断結果を字種関連素性として取得する。なお、素性取得部１５は、例えば、編集箇所が「Ａ」と「ａ」の用語対に対して合致すると判断し、編集箇所が「Ａ」と「ｂ」の用語対に対して合致しないと判断する。

【００５０】

素性取得部１５は、例えば、用語対格納部１１の用語対ごとに、機械学習部１６が利用する教師あり機械学習法とは異なる分類方法、または同一の分類方法であるが学習データが異なる分類方法により、用語対が異表記の用語対であるか否かを判断し、判断結果を辞書関連素性として取得する。かかる辞書関連素性を利用する機械学習法を、スタッキングアルゴリズムによる方法という。機械学習部１６が利用する教師あり機械学習法とは異なる分類方法とは、上記の機械学習法がＳＶＭである場合、ＳＶＭとは異なる決定木などの機械学習法、後述するルールに基づく分類方法等である。

【００５１】

スタッキングアルゴリズムは、詳細には、例えば、以下の手順による分類方法である。まず、ＪＵＭＡＮ辞書を使って教師データを作成する。つまり、ＪＵＭＡＮ辞書の単語の集合から、編集距離が１文字の単語対を取り出す。ここで、編集距離が１文字の単語対は、９０４６１２組、取り出せる。そのうち、代表表記が等しい単語対（２５９３４組）を取り出す。次に、ＪＵＭＡＮ辞書で、代表表記が等しい単語対を正例、そうでないものを負例とする。以上により、教師データを作成する。

【００５２】

次に、その教師データを学習データとした機械学習を行う。なお、教師データは、上述した教師データに限らず、他の教師データを用いてもよい。また、機械学習の際に利用する素性は、本発明の全素性（Ｓ１からＳ６８の素性）のうち、Ｓ５４の素性を取り除いた素性を利用する。なお、機械学習の際に利用する素性は、他の素性を用いてもよい。

【００５３】

そして、実際に、Ｓ５４の素性を付与したいデータを、上記学習結果を利用して、分類する。分類結果において正例となったか、負例となったかの情報をＳ５４の素性として、そのデータに付与する。

【００５４】

そして、Ｓ５４の素性が付与された学習データ（６８の素性を有する）を用いて、問題となる用語対に対して、機械学習を行うことで、問題となる用語対が異表記対であるか否かを判断していく。

【００５５】

スタッキングアルゴリズムによる方法では、ＪＵＭＡＮ辞書で、代表表記が一致するか否かについて学習した結果を素性として付与できるので、実際にＪＵＭＡＮ辞書に記載されていない用語対に対しても、ＪＵＭＡＮ辞書で、代表表記が一致するとされる傾向のある用語対か否かの情報を付与できることとなる。

【００５６】

素性取得部１５は、例えば、用語対格納部１１の用語対ごとに、用語対が有する２つの用語の代表表記を、用語辞書１３から取得し、取得した２つの代表表記が同一であるか否かを判断し、判断結果を辞書関連素性として取得する。

【００５７】

素性取得部１５は、例えば、用語対格納部１１の用語対ごとに、用語辞書１３から２つの用語の読みを取得し、２つの用語の読みが一致するか否かを判断し、判断結果を辞書関連素性として取得する。また、素性取得部１５は、例えば、用語対格納部１１の用語対ごとに、用語対が有する２つの用語の文字数が同数であり、かつ、用語辞書１３から２つの用語の読みを取得し、２つの用語の読みが一致するか否かを判断し、判断結果を辞書関連

素性として取得しても良い。

【００５８】

なお、上述した判断結果とは、例えば、上記条件に合致する場合の判断結果は「１」、その他の場合の判断結果は「０」などである。

【００５９】

差分文字取得手段１５１は、編集距離が２つの用語対について、２つの差分文字の組を、それぞれ取得する。例えば、編集距離が２つの用語対が、（１）「できる」「出来る」（２）「理解できる」「できる」（３）「ＩＸ（ローマ数字の９）」「９」である場合を考える。（１）は両方の用語対が同じ文字数である場合である。（２）はどちらか一方の用語の文字数がもう片方の用語の文字数より２つ多いまたは、少ない場合である。（３）はどちらか一方の用語の文字数がもう片方の用語の文字数より１つ多いまたは、少ない場合である。（１）の場合、差分文字取得手段１５１は、「できる」および「出来る」の用語に対して、前方から後方に１，２，３・・・と文字に番号をつけ、それぞれの用語で同じ文字番号を持ち、違う文字である「で」「出」と「き」「来」が差分文字であるとして、「で」「出」と「き」「来」の２組の差分文字の組を取得する。（２）の場合、差分文字取得手段１５１は、「理」「」と「解」「」（「」はＮＵＬＬである）の２組の差分文字の組を取得する。（３）の場合、差分文字取得手段１５１は、「Ｉ」「９」と「Ｘ」「」の２組の差分文字、または「Ｉ」「」と「Ｘ」「９」の２組の差分文字を取得する。

【００６０】

また、差分文字取得手段１５１は、編集距離が１の用語対について、差分文字の組を１組取得する。例えば、編集距離が１の用語対が、（１）「ご苦労」「御苦労」（２）「Ｆｉｒｅｆｏｘ」「ＦｉｒｅＦｏｘ」（３）「肝炎ウイルス」「肝炎ウィルス」（４）「文学史上」「文学史」（５）「咲き分け」「咲分け」（６）「クロゼット」「クローゼット」（７）「大人・子供」「大人子供」（８）「第１位」「第一位」である場合、差分文字取得手段１５１は、それぞれ（１）「ご」「御」（２）「ｆ」「Ｆ」（３）「イ」「ィ」（４）「上」「」（５）「き」「」（６）「」「ー」（７）「・」「」（８）「１」「一」を取得する。

【００６１】

さらに、差分文字取得手段１５１は、編集距離が３以上の用語対について、３組以上の差分文字の組を取得する。例えば、編集距離が４の用語対が、「１０２５位」「千二十五位」である場合、差分文字取得手段１５１は、「１」「千」、「０」「二」、「２」「十」「５」「五」という４組の差分文字を取得する。ここで、差分文字とは、２つの用語の異なる文字である。

【００６２】

素性取得手段１５２は、差分文字取得手段１５１が取得した２つの差分文字を、独立に対象として、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を、２組取得する。例えば、編集距離が２つの用語対が、（１）「できる」「出来る」（２）「理解できる」「できる」（３）「ＩＸ」「９」である場合を考える。（１）の用語対について、素性取得手段１５２は、「で」「出」と「き」「来」の２組の差分文字の組のそれぞれを対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、２種類のテストデータを作成する。素性取得手段１５２は、例えば、用語対が有する２つの用語の編集箇所の字種が異なり、かつ、２つの用語の編集箇所が同じ値の数字であるか否かを示す字種関連素性について、「で」「出」の編集箇所が同じ値の数字でないと判断し、当該字種関連素性「０」を取得する。また、素性取得手段１５２は、例えば、用語辞書１３から２つの用語の読みを取得し、２つの用語の読みが一致するか否かを示す辞書関連素性「１」を取得する。素性取得手段１５２は、用語辞書１３から「出」の読み「で」を取得し、「で」と「出」の読みが一致すると判断する。また、素性取得手段１５２は、例えば、差分文字「で」「出」に対して、差分文字（編集箇所）の前後の文字という素性について、前の文字の素性「」（なし）、後の文字の素性「き」と「来」を取得する。また素性取得手段１５２は、例えば、差分文字「き」「来」に対して、差分文字の

前後の文字という素性について、前の文字の素性「出」と「で」、後の文字の素性「る」を取得する。かかる処理により、別の差分文字も素性に含めることとなる。

【００６３】

また、（２）の用語対について、素性取得手段１５２は、（１）と同様に、「理」「」と「解」「」の２組の差分文字の組のそれぞれを対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、２種類のテストデータを作成する。

【００６４】

さらに、（３）の用語対について、素性取得手段１５２は、（１）（２）と同様に、例えば、「Ｉ」「９」と「Ｘ」「」の２組の差分文字の組のそれぞれを対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、２種類のテストデータを作成する。

【００６５】

また、素性取得手段１５２は、差分文字取得手段１５１が取得した１組以上の差分文字を用いて、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する。なお、字種関連素性、辞書関連素性、類似度素性などの素性を取得する具体的な方法は後述する。

【００６６】

機械学習部１６は、各用語対に対して、学習データ格納部１２の２以上の学習データと、素性取得部１５が取得した複数の素性とを用いて、教師あり機械学習法により、用語対格納部１１の各用語対が異表記の用語対であるか否かを判断する。

【００６７】

機械学習部１６は、用語対格納部１１の各用語対が異表記の用語対であるか否かを判断するとともに、異表記の用語対である確度を示すスコアも取得しても良い。

【００６８】

機械学習部１６は、素性取得手段１５２が取得した２組の複数の素性のうちの組ごとに、各組の複数の素性と、学習データ格納部１２の２以上の学習データとを用いて、教師あり機械学習法により、用語対格納部１１の各組の複数の素性が異表記の用語対に対応する素性の組であるか否かを判断し、２つの判断結果を用いて、編集距離が２である用語対が異表記の用語対であるか否かを判断する。

【００６９】

教師あり機械学習法のアルゴリズムは問わない。教師あり機械学習法とは、例えば、サポートベクターマシン（ＳＶＭ）などである。ＳＶＭは、「http://chasen.org/~taku/software/TinySVM/」「http://ja.wikipedia.org/wiki/%E3%82%B5%E3%83%9D%E3%83%BC%E3%83%88%E3%83%99%E3%82%AF%E3%82%BF%E3%83%BC%E3%83%9E%E3%82%B7%E3%83%B3」（平成２１年１２月１２日検索）などに記載されている。なお、教師あり機械学習法の詳細は、後述する。

【００７０】

また、上記の、２つの判断結果を用いてとは、２つとも正例とされた場合に異表記の用語対としても良いし、２つとも負例とされた場合に異表記の用語対ではないとしても良いし、２つのスコアのうちのスコアが０に近い方のスコアを採用して、採用したスコアが正の場合は正例（異表記の用語）、負の場合は負例（異表記の用語でない）と判断しても良いし、スコアの絶対値が大きい方のスコアを採用して、採用したスコアが正の場合は正例（異表記の用語）、負の場合は負例（異表記の用語でない）と判断しても良い。また、２つのスコアのうち、小さい方のスコアを取得し、当該小さい方のスコアが正の場合は正例、負の場合は負例と判断しても良い。つまり、２つの判断結果の用い方は問わない。なお、上記の（２）の場合（どちらか一方の用語の文字数がもう片方の用語の文字数より２つ多いまたは、少ない場合）、大規模類似語リストの中から、約１万５千のタグ付けを行った結果、このパターンの２文字差分データには、異表記対であると判定する用語対はなかった。

【００７１】

さらに、２組の差分文字の組（例えば、「Ｉ」「９」と「Ｘ」「」、または「Ｉ」「」と「Ｘ」「９」）、つまり２つの問題（問題１、問題２）ができる場合、それぞれの差分文字を対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、４種類のテストデータを作成する。そして、２つの問題ごとに、算出したスコアが０に近い方を取得し、問題ごとのスコアのうちの、絶対値が高いスコアを当該問題のスコアとし、スコアが正の場合は正例、負の場合は負例と判断しても良い。例えば、編集距離が２の用語対が（３）「ＩＸ」「９」である場合、問題「Ｉ」「９」と「Ｘ」「」、および「Ｉ」「」と「Ｘ」ができる。そして、機械学習部１６は、「Ｉ」「９」と「Ｘ」「」のスコアの小さい方を取得し、また、「Ｉ」「」と「Ｘ」「９」のスコアの小さい方を取得し、２つの取得されたスコアのうち、値が大きい方を「ＩＸ」「９」の用語対におけるスコアとする。そして、機械学習部１６は、当該スコアが正の場合は正例、負の場合は負例と判断しても良い。なお、例えば、機械学習部１６は、「Ｉ」「９」と「Ｘ」「」のスコアが０に近い方を取得し、また、「Ｉ」「」と「Ｘ」「９」のスコアが０に近い方を取得し、２つの取得されたスコアのうち、絶対値が大きい方を「ＩＸ」「９」の用語対におけるスコアとしても良い。つまり、４種類のテストデータの判断結果を如何に用いてスコアを算出するかは問わない。

【００７２】

出力部１７は、機械学習部１６における判断結果を出力する。また、出力部１７は、機械学習部１６が取得したスコアを出力しても良い。判断結果とは、各用語対が異表記の用語対であるか否かを示す情報、または異表記の１以上の用語対、または異表記でない１以上の用語対などである。また、出力部１７は、判断結果とスコアの両方を出力しても良いし、一方を出力しても良い。

【００７３】

また、出力とは、ディスプレイへの表示、プロジェクターを用いた投影、プリンタへの印字、音出力、外部の装置への送信、記録媒体への蓄積、他の処理装置や他のプログラムなどへの処理結果の引渡しなどを含む概念である。

【００７４】

閾値格納手段１７１は、スコアの閾値を格納している。

【００７５】

閾値判断手段１７２は、機械学習部１６が取得したスコアが閾値以上または閾値より大きいか否かを判断する。

【００７６】

出力手段１７３は、閾値判断手段１７２が閾値以上または閾値より大きいと判断したスコアに対応する用語対を、異表記の用語対であるとの判断結果とし、判断結果または異表記の用語対または異表記でない用語対のいずれか１以上を出力する。

【００７７】

用語対格納部１１、学習データ格納部１２、用語辞書１３、および閾値格納手段１７１は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。

【００７８】

用語対格納部１１、学習データ格納部１２、および用語辞書１３に格納されている情報が記憶される過程は問わない。

【００７９】

素性取得部１５、機械学習部１６、閾値判断手段１７２は、通常、ＭＰＵやメモリ等から実現され得る。素性取得部１５等の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。

【００８０】

出力部１７は、ディスプレイやスピーカー等の出力デバイスを含むと考えても含まないと考えても良い。出力部１７は、出力デバイスのドライバーソフトまたは、出力デバイスのドライバーソフトと出力デバイス等で実現され得る。

【００８１】
次に、異表記取得装置１の動作について図２のフローチャートを用いて説明する。ここでは、異表記取得装置１は、編集距離が１の用語対に対して、異表記の用語対であるか否かを判断することとする。
【００８２】
（ステップＳ２０１）受付部１４は、動作開始の指示を受け付けたか否かを判断する。指示を受け付ければステップＳ２０２に行き、受け付けなければステップＳ２０１に戻る。
【００８３】
（ステップＳ２０２）素性取得部１５は、カウンタｉに１を代入する。
【００８４】
（ステップＳ２０３）素性取得部１５は、ｉ番目の用語対が用語対格納部１１に存在するか否かを判断する。ｉ番目の用語対が存在すればステップＳ２０４に行き、存在しなければ処理を終了する。
【００８５】
（ステップＳ２０４）素性取得部１５は、用語対格納部１１から、ｉ番目の用語対を読み出す。
【００８６】
（ステップＳ２０５）素性取得部１５は、ｉ番目の用語対の素性を取得する処理を行う。素性取得処理について、図３のフローチャートを用いて説明する。
【００８７】
（ステップＳ２０６）機械学習部１６は、ステップＳ２０５で取得された複数の素性と、学習データ格納部１２の２以上の学習データとを用いて、教師あり機械学習を行い、スコアを取得する。
【００８８】
（ステップＳ２０７）出力部１７は、ステップＳ２０６で取得されたスコアを用いて、ｉ番目の用語対は異表記の用語対であるか否かを判断する。例えば、出力部１７を構成する閾値判断手段１７２は、閾値格納手段１７１から閾値を読み出し、ステップＳ２０６で取得されたスコアが閾値より大きいまたは閾値以上であれば、ｉ番目の用語対は異表記の用語対であると判断し、スコアが閾値以下または閾値より小さい場合は、ｉ番目の用語対は異表記の用語対でない、と判断する。
【００８９】
（ステップＳ２０８）出力部１７は、ステップＳ２０８での判断結果が、異表記の用語対であればステップＳ２０９に行き、異表記の用語対でなければステップＳ２１０に行く。
【００９０】
（ステップＳ２０９）出力部１７は、ｉ番目の用語対を異表記の用語対であるとして出力する。
【００９１】
（ステップＳ２１０）素性取得部１５は、カウンタｉを１，インクリメントする。ステップＳ２０３に戻る。
【００９２】
次に、ステップＳ２０５の素性取得処理について、図３のフローチャートを用いて説明する。
【００９３】
（ステップＳ３０１）素性取得部１５を構成する差分文字取得手段１５１は、２つの用語の編集箇所を取得する。
【００９４】
（ステップＳ３０２）素性取得部１５の素性取得手段１５２は、ステップＳ３０１で取得された編集箇所を用いて、字種関連素性を取得する。字種関連素性の具体的な取得方法

については後述する。

【００９５】

（ステップＳ３０３）素性取得手段１５２は、用語辞書１３を用いて、辞書関連素性を取得する。辞書関連素性の具体的な取得方法については後述する。

【００９６】

（ステップＳ３０４）素性取得手段１５２は、２つの用語の類似度を取得する。この類似度は、類似度素性である。

【００９７】

（ステップＳ３０５）素性取得手段１５２は、その他、予め決められた素性を取得する。その他の予め決められた素性の例は、後述する。

【００９８】

（ステップＳ３０６）素性取得手段１５２は、スタッキングアルゴリズムを使用して、ステップＳ３０２からステップＳ３０５において取得した複数の素性を用いて、ステップＳ２０６における教師あり機械学習法とは異なる分類方法により、用語対が異表記の用語対であるか否かを判断し、その判断結果を取得する。

【００９９】

以下、本実施の形態における異表記取得装置１の具体的な動作について説明する。

【０１００】

今、用語辞書１３は、例えば、図４に示すような構造を有する、とする。図４において、一用語の情報は、一レコードになっている。各レコードは、「用語」「読み」「品詞」「代表表記」「カテゴリ」「ドメイン」の属性値を有する。用語辞書１３は、例えば、ＪＵＭＡＮ辞書（「http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html」参照［平成２１年１２月１３日検索］）である。また、用語辞書１３は、例えば、日本語ワードネット辞書（http://nlpwww.nict.go.jp/wn-ja/index.ja.html参照［平成２１年１２月１３日検索］）や、異体字辞書や、ＥＤＲ電子化辞書（http://www2.nict.go.jp/r/r312/EDR/J_index.html参照［平成２１年１２月１３日検索］）である。なお、異体字辞書とは、異体字の対を有する辞書である。異体字とは、読み方や用い方が同じでも字形に異なる部分のある字体のことである。旧字体と新字体がある漢字などに多く見られ、例えば「沢」と「澤」は異体字の関係にある。なお、異体字辞書は、異体字ではないが、異体字のように代替可能な漢字の対を有しても良い。さらに、用語辞書１３は、異体字辞書とは別に、異体字のように代替可能な漢字の対を有する辞書を有しても良い。

【０１０１】

また、学習データ格納部１２に格納されている学習データが有する複数の素性、および素性取得部１５が取得する複数の素性は、ここでは、６８種類である、とする。以下に、６８の素性（Ｓ１からＳ６８）について説明する。また、以下、用語対の具体例として、用語対「ショウウインドウ」「ショーウインドウ」を用いて、素性を例示する。

【０１０２】

Ｓ１は、「一つ目の表記の編集箇所」であり、上記具体例では、「ウ」である。素性Ｓ１を取得する場合、差分文字取得手段１５１は、用語対の構成する２つの用語を１文字ずつずらしながら文字を比較し、編集箇所を得る。例えば、差分文字取得手段１５１は、「ショウウインドウ」の１文字目「シ」と、「ショーウインドウ」の１文字目「シ」とから比較し、同一と判断し、２文字目も同一と判断し、３文字目「ウ」と「ー」とが異なると判断し、一つ目の表記の編集箇所「ウ」と二つ目の表記の編集箇所「ー」とを取得する。

【０１０３】

Ｓ２は、「二つ目の表記の編集箇所」であり、上記具体例では、「ー」である。一つ目の表記とは用語対を構成する一つ目の用語（例えば、「ショウウインドウ」）であり、二つ目の表記とは用語対を構成する二つ目の用語（例えば、「ショーウインドウ」）である。

【０１０４】

Ｓ３は「編集箇所の前方の１文字」であり、上記具体例では、「ョ」である。

【０１０５】
Ｓ４は「編集箇所の後方の１文字」であり、上記具体例では、「ウ」である。
【０１０６】
Ｓ５は、「編集箇所の前方の連続する２文字」であり、上記具体例では、「ショ」である。
【０１０７】
Ｓ６は、「編集箇所の前方の連続する３文字」であり、上記具体例では、「ショ」である。
【０１０８】
Ｓ７は、「編集箇所の前方２文字目の文字」であり、上記具体例では、「シ」である。
【０１０９】
Ｓ８は「編集箇所の前方３文字目の文字」であり、上記具体例では、「(del)」である。(del)とは、文字が無いことを示す。
【０１１０】
Ｓ９は「編集箇所の後方の２文字」であり、上記具体例では、「ウイ」である。
【０１１１】
Ｓ１０は「編集箇所の後方の３文字」であり、上記具体例では、「ウイン」である。
【０１１２】
Ｓ１１は「編集箇所の後方２文字目の文字」であり、上記具体例では、「イ」である。
【０１１３】
Ｓ１２は「編集箇所の後方３文字目の文字」であり、上記具体例では、「ン」である。
【０１１４】
Ｓ１３は「'Ｓ１の情報－Ｓ２の情報'とした文字列」であり、上記具体例では、「ウ－ー」である。
【０１１５】
Ｓ１４は「'Ｓ３の情報－Ｓ１３の情報'とした文字列」であり、上記具体例では、「ョ－ウ－ー」である。
【０１１６】
Ｓ１５は「'Ｓ５の情報－Ｓ１３の情報'とした文字列」であり、上記具体例では、「ショ－ウ－ー」である。
【０１１７】
Ｓ１６は「'Ｓ６の情報－Ｓ１３の情報'とした文字列」であり、上記具体例では、「ショ－ウ－ー」である。
【０１１８】
Ｓ１７は「'Ｓ１３の情報－Ｓ４の情報'」であり、上記具体例では、「ウ－－－ウ」である。
【０１１９】
Ｓ１８は「'Ｓ３の情報－Ｓ１３の情報－Ｓ４の情報'」であり、上記具体例では、「ョ－ウ－－－ウ」である。
【０１２０】
Ｓ１９は「'Ｓ５の情報－Ｓ１３の情報－Ｓ４の情報'とした文字列」であり、上記具体例では、「ショ－ウ－－－ウ」である。
【０１２１】
Ｓ２０は「'Ｓ６の情報－Ｓ１３の情報－Ｓ４の情報'とした文字列」であり、上記具体例では、「ショ－ウ－－－ウ」である。
【０１２２】
Ｓ２１は「'Ｓ１３の情報－Ｓ７の情報'とした文字列」であり、上記具体例では、「ウ－－－シ」である。
【０１２３】
Ｓ２２は「'Ｓ３の情報－Ｓ１３の情報－Ｓ７の情報'とした文字列」であり、上記具体

例では、「ョーウーーーシ」である。
【０１２４】
Ｓ２３は「'Ｓ５の情報－Ｓ１３の情報－Ｓ７の情報'とした文字列」であり、上記具体例では、「ショーウーーーシ」である。
【０１２５】
Ｓ２４は「'Ｓ６の情報－Ｓ１３の情報－Ｓ７の情報'とした文字列」であり、上記具体例では、「ショーウーーーシ」である。
【０１２６】
Ｓ２５は「'Ｓ１３の情報－Ｓ８の情報'とした文字列」であり、上記具体例では、「ウーーー（ｄｅｌ）」である。
【０１２７】
Ｓ２６は「'Ｓ３の情報－Ｓ１３の情報－Ｓ８の情報'とした文字列」であり、上記具体例では、「ョーウーーー（ｄｅｌ）」である。
【０１２８】
Ｓ２７は「'Ｓ５の情報－Ｓ１３の情報－Ｓ８の情報'とした文字列」であり、上記具体例では、「ショーウーーー（ｄｅｌ）」である。
【０１２９】
Ｓ２８は「'Ｓ６　の情報－Ｓ１３の情報－Ｓ８の情報'とした文字列」であり、上記具体例では、「ショーウーーー（ｄｅｌ）」である。なお、２つの用語が与えられ、編集箇所が判断できれば、単なる文字列の処理（操作）により、素性取得手段１５２は、Ｓ３からＳ２８の素性を取得できる。
【０１３０】
Ｓ２９は「Ｓ１の字種」であり、上記具体例では、「カタカナ」である。文字を与えられた場合、当該文字の字種（漢字、ひらがな、かたかな、アルファベット等）を取得する技術は公知技術である。
【０１３１】
Ｓ３０は「Ｓ２の字種」であり、上記具体例では、「カタカナ」である。
【０１３２】
Ｓ３１は「Ｓ３の字種」であり、上記具体例では、「カタカナ」である。
【０１３３】
Ｓ３２は「Ｓ４の字種」であり、上記具体例では、「カタカナ」である。
【０１３４】
Ｓ３３は「Ｓ１３の字種」であり、上記具体例では、「カタカナ」である。
【０１３５】
Ｓ３４は「Ｓ１４の字種」であり、上記具体例では、「カタカナ」である。
【０１３６】
Ｓ３５は「Ｓ１７の字種」であり、上記具体例では、「カタカナ」である。
【０１３７】
Ｓ３６は「Ｓ１８の字種」であり、上記具体例では、「カタカナ」である。
【０１３８】
Ｓ３７は「Ｓ１の品詞」であり、上記具体例では、「名詞」である。ここで、文字の品詞は、その文字（ここではＳ１）が属している用語の品詞である。例えば、用語に対して、形態素解析をかけ、用語を単語に区切り、品詞情報を取得する。そして、文字の品詞は、当該取得した品詞情報が示す品詞である。
【０１３９】
Ｓ３８は「Ｓ２の品詞」であり、上記具体例では、「名詞」である。
【０１４０】
Ｓ３９は「Ｓ３の品詞」であり、上記具体例では、「名詞」である。
【０１４１】
Ｓ４０は「Ｓ４の品詞」であり、上記具体例では、「名詞」である。

【０１４２】
Ｓ４１は「Ｓ１３の品詞」であり、上記具体例では、「名詞」である。
【０１４３】
Ｓ４２は「Ｓ１４の品詞」であり、上記具体例では、「名詞」である。
【０１４４】
Ｓ４３は「Ｓ１７の品詞」であり、上記具体例では、「名詞」である。
【０１４５】
Ｓ４４は「Ｓ１８の品詞」であり、上記具体例では、「名詞」である。
【０１４６】
Ｓ４５は「Ｓ１の品詞と位置情報」であり、上記具体例では、「名詞，３」である。ここで「３」は、３文字目であることを示す。
【０１４７】
Ｓ４６は「Ｓ２の品詞と位置情報」であり、上記具体例では、「名詞，３」である。
【０１４８】
Ｓ４７は「Ｓ３の品詞と位置情報」であり、上記具体例では、「名詞，２」である。
【０１４９】
Ｓ４８は「Ｓ４の品詞と位置情報」であり、上記具体例では、「名詞，４」である。
【０１５０】
Ｓ４９は「Ｓ１３の品詞と位置情報」であり、上記具体例では、「名詞，３」である。
【０１５１】
Ｓ５０は「Ｓ１４の品詞と位置情報」であり、上記具体例では、「名詞，２」である。
【０１５２】
Ｓ５１は「Ｓ１７の品詞と位置情報」であり、上記具体例では、「名詞，６」である。
【０１５３】
Ｓ５２は「Ｓ１８の品詞と位置情報」であり、上記具体例では、「名詞，３」である。
【０１５４】
Ｓ５３は「日本語用語対の類似度」であり、上記具体例では、例えば、０．９である。
【０１５５】
Ｓ５４は「スタッキングアルゴリズムを使用して、日本語用語対のＪＵＭＡＮ辞書の代表表記が一致するかどうか」を示す情報であり、上記具体例では、「１」である。つまり、ここでは、機械学習部１６が利用する教師あり機械学習法とは異なる分類方法は、用語対を構成する２つの用語の、ＪＵＭＡＮ辞書における代表表記が一致するか否かにより分類する以下の方法である。まず、ＪＵＭＡＮ辞書の単語の集合から、編集距離が１文字の単語対を取り出す。ここで、編集距離が１文字の単語対は、９０４６１２組、取り出せる。そのうち、代表表記が等しい単語対（２５９３４組）を取り出す。次に、ＪＵＭＡＮ辞書で、代表表記が等しい単語対を正例、そうでないものを負例とする。以上により、教師データを作成する。次に、その教師データを学習データとした機械学習を行う。なお、教師データは、上述した教師データに限らず、他の教師データを用いてもよい。また、機械学習の際に利用する素性は、本発明の全素性（Ｓ１からＳ６８の素性）のうち、Ｓ５４の素性を取り除いた素性を利用する。なお、機械学習の際に利用する素性は、他の素性を用いてもよい。そして、実際に、Ｓ５４の素性を付与したいデータを、上記学習結果を利用して、分類する。分類結果において正例となったか、負例となったかの情報をＳ５４の素性として、そのデータに付与する。スタッキングアルゴリズムによる方法では、ＪＵＭＡＮ辞書で、代表表記が一致するか否かについて学習した結果を素性として付与できるので、実際にＪＵＭＡＮ辞書に記載されていない用語対に対しても、ＪＵＭＡＮ辞書で、代表表記が一致するとされる傾向のある用語対か否かの情報を付与できることとなる。
【０１５６】
Ｓ５５は「日本語用語対の文字数が同数で編集箇所が両方とも数字の場合であり、同じ値か違う値かどうか」であり、上記具体例では、「０」である。なお、「２次キャッシュ」と「二次キャッシュ」の用語対の場合、「一週間あたり」「１週間あたり」の用語対の

場合は、Ｓ５５の素性は「１」となる。なお、文字数が同数である条件をはずし、Ｓ５５は、「日本語用語対の編集箇所が両方とも数字の場合であり、同じ値か違う値かどうか」が好適である。

【０１５７】

Ｓ５６は「日本語用語対の文字数が同数で編集箇所が両方ともひらがなの場合であり、同じ音声か違う音声かどうか」であり、上記具体例では、「０」である。なお、「おかぁちゃん」「おかあちゃん」の用語対の場合は、Ｓ５６の素性は「１」となる。なお、文字数が同数である条件をはずし、Ｓ５６は、「日本語用語対の編集箇所が両方ともひらがなの場合であり、同じ音声か違う音声かどうか」が好適である。

【０１５８】

Ｓ５７は「日本語用語対の文字数が同数で編集箇所が両方ともカタカナの場合であり、同じ音声か違う音声かどうか」であり、上記具体例では、「１」である。なお、「オリーブ・オイル」「オリーヴ・オイル」の用語対の場合、「ウインドウ」「ウィンドウ」の用語対の場合も、Ｓ５７の素性は「１」となる。なお、文字数が同数である条件をはずし、Ｓ５７は、「日本語用語対の編集箇所が両方ともカタカナの場合であり、同じ音声か違う音声かどうか」が好適である。

【０１５９】

Ｓ５８は「日本語用語対の文字数が同数で編集箇所が両方ともローマ字の場合であり、大文字と小文字の違いだけかどうか」であり、上記具体例では、「０」である。なお、「３００ｋｂｐｓ」「３００Ｋｂｐｓ」の用語対の場合、「Ｗｉｎｄｏｗｓ上」「ｗｉｎｄｏｗｓ上」の用語対の場合は、Ｓ５８の素性は「１」となる。なお、文字数が同数である条件をはずし、Ｓ５８は、「日本語用語対の編集箇所が両方ともローマ字の場合であり、大文字と小文字の違いだけかどうか」が好適である。また、「Ｗｉｎｄｏｗｓ」は登録商標です。

【０１６０】

Ｓ５９は「日本語用語対の文字数が同数で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かどうか」であり、上記具体例では、「０」である。なお、「触れるくらい」「触れるぐらい」の用語対の場合、「飲むぐらい」「飲むくらい」の用語対の場合は、Ｓ５９の素性は「１」となる。なお、Ｓ５９は、文字数が同数である条件をはずし、「日本語用語対の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かどうか」が好適である。

【０１６１】

Ｓ６０は「日本語用語対の文字数が同数で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かどうか」であり、上記具体例では、「０」である。なお、Ｓ６０は、文字数が同数である条件をはずし、「日本語用語対の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かどうか」が好適である。

【０１６２】

Ｓ６１は「編集箇所が日本語用語対の一方にしかなく、その編集箇所が'化'、'系'、'類'、'型'、'形'、'氏'、'ー'、'・'かどうか」であり、上記具体例では、「０」である。なお、「サーバ」「サーバー」の用語対の場合、「ハンセン病患者」「ハンセン氏病患者」の用語対の場合、「日本語パッチ」「日本語化パッチ」の用語対の場合、「３０種類ほど」「３０種ほど」の用語対の場合は、Ｓ６１の素性は「１」となる。

【０１６３】

Ｓ６２は「編集箇所が日本語用語対の一方にしかなく、その編集箇所の用語が日本語用語対の最後の文字と一致するかどうか」であり、上記具体例では、「０」である。なお、「妊娠・授乳中」「妊娠中・授乳中」の用語対の場合、「国産・輸入車」「国産車・輸入車」の用語対の場合は、Ｓ６２の素性は「１」となる。

【０１６４】

Ｓ６３は「編集箇所が日本語用語対の一方にしかなく、その編集箇所が桁数をあらわす用語かどうか（例えば、"千""万"など）」であり、上記具体例では、「０」である。なお

、「２万５０００人」「２５０００人」の用語対の場合、「１万６５００円」「１６５００円」の用語対の場合は、Ｓ６３の素性は「１」となる。

【０１６５】

Ｓ６４は「日本語用語対のＪＵＭＡＮ辞書の定義されている代表表記が一致するかどうか」であり、上記具体例（用語対「ショウウインドウ」「ショーウインドウ」）では、例えば、「１」である。なお、素性取得手段１５２は、用語対を構成する各用語の代表表記を、用語辞書１３から取得し、比較することにより、素性を取得する。

【０１６６】

Ｓ６５は「日本語用語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どうか」であり、上記具体例（用語対「ショウウインドウ」「ショーウインドウ」）では、例えば、「１」である。素性取得手段１５２は、用語対を構成する各用語をキーとして、日本語ワードネット辞書を検索し、類義語対として定義されているか否かを判断する。本処理は、通常の検索処理である。

【０１６７】

Ｓ６６は「日本語用語対の編集箇所が異体字辞書に異体字として定義されているかどうか」であり、上記具体例（用語対「ショウウインドウ」「ショーウインドウ」）では、例えば、「０」である。異体字辞書は、２つの異体字の対の情報を有する。

【０１６８】

Ｓ６７は「日本語用語対の文字数が同数で編集箇所が漢字とひらがなの場合であり、ＪＵＭＡＮ辞書の読みが一致するかどうか」であり、上記具体例（用語対「ショウウインドウ」「ショーウインドウ」）では、「０」である。

【０１６９】

Ｓ６８は「日本語用語対の文字数が同数で編集箇所が両方とも漢字の場合であり、ＪＵＭＡＮ辞書の読みが一致するかどうか」であり、上記具体例（用語対「ショウウインドウ」「ショーウインドウ」）では、「０」である。

【０１７０】

また、上記の６８の素性は、上述したように、字種関連素性、辞書関連素性、類似度素性、編集箇所文字素性などが含まれる。

【０１７１】

また、上記の６８の素性をグループ化すると、例えば、以下のＧ１からＧ７のグループに分かれる、と考えられる。

【０１７２】

Ｇ１は、Ｓ１からＳ５２の素性であり、編集箇所とその周辺の文字列に関する情報である編集箇所文字素性である。

【０１７３】

Ｇ２は、Ｓ５３の素性であり、類似度素性である。

【０１７４】

Ｇ３は、Ｓ５４の素性であり、スタッキングアルゴリズムを使用した情報である素性である。

【０１７５】

Ｇ４は、Ｓ５５からＳ６０の素性であり、編集箇所に関する情報である編集箇所関連素性である。

【０１７６】

Ｇ５は、Ｓ６１からＳ６３の素性であり、用語対のパターンに関する情報である用語対パターン素性である。

【０１７７】

Ｇ６は、Ｓ６４からＳ６６の素性であり、種々の辞書による情報である辞書関連素性である。

【０１７８】

Ｇ７は、Ｓ６７からＳ６８の素性であり、読みに関する情報である読み関連素性である

。
【０１７９】
そして、まず、異表記取得装置１において、予め正しい異表記の用語対のデータ（正例）を人手で構築しておき、正例の用語対と、正例であることを示す正負情報（例えば、「１」）とを対応付けて、学習データ格納部１２に格納しておく。また、異表記取得装置１において、予め異表記でない用語対のデータ（負例）を人手で構築しておき、負例の用語対と、負例であることを示す正負情報（例えば、「０」）とを対応付けて、学習データ格納部１２に格納しておく。
【０１８０】
次に、異表記取得装置１の素性取得部１５により、各用語対の、上述した６８の素性を取得し、正負情報または用語対と対応付けて、６８の素性を学習データ格納部１２に蓄積する。
【０１８１】
以上の処理により、学習データ格納部１２の学習データが構築された。
【０１８２】
次に、異表記の用語対であるか否かを判断したい１以上の用語対を用語対格納部１１に格納する。
【０１８３】
そして、ユーザは、異表記取得装置１に、動作開始の指示を入力する。すると、受付部１４は、動作開始の指示を受け付ける。
【０１８４】
次に、用語対格納部１１に格納されている用語対を順に、以下のように処理する。つまり、素性取得部１５は、各用語対の素性を取得する処理を行う。かかる素性取得処理については説明済みである。
【０１８５】
次に、機械学習部１６は、取得された６８の素性と、学習データ格納部１２の学習データとを用いて、教師あり機械学習を行い、スコアを取得する。
【０１８６】
次に、出力部１７は、取得された各用語対のスコアを用いて、各用語対は異表記の用語対であるか否かを判断する。
【０１８７】
次に、出力部１７は、異表記の用語対であると判断した用語対のみ出力する。ここで、出力とは、たとえば、予め決められた記憶媒体への蓄積である。
（機械学習について）
【０１８８】
以下、機械学習部１６が行う機械学習、および機械学習部１６が行う教師あり機械学習法とは異なる分類方法（スタッキングアルゴリズムで利用）について説明する。
【０１８９】
まず、機械学習法とは、問題-解の組のセットを多く用意し、それで学習を行ない、どういう問題のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときも解を推測できるようにする方法である(例えば、非特許文献２～非特許文献４参照)。
【０１９０】
どういう問題のときにどういう解になるかという、問題の状況を機械に伝える際に、素性(解析に用いる情報で問題を構成する各要素)が必要になる。問題を素性によって表現するのである。
【０１９１】
すなわち、機械学習の手法は、素性の集合-解の組のセットを多く用意し、それで学習を行ない、どういう素性の集合のときにどういう解になるかを学習し、その学習結果を利用して、新しい問題のときもその問題から素性の集合を取り出し、その素性の場合の解を

推測する方法である。

【０１９２】

機械学習の手法として、例えば、ｋ近傍法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法などの手法を用いることができる。

【０１９３】

ｋ近傍法は、最も類似する一つの事例のかわりに、最も類似するｋ個の事例を用いて、このｋ個の事例での多数決によって分類先（解）を求める手法である。ｋは、あらかじめ定める整数の数字であって、一般的に、１から９の間の奇数を用いる。

【０１９４】

シンプルベイズ法は、ベイズの定理にもとづいて各分類になる確率を推定し、その確率値が最も大きい分類を求める分類先とする方法である。

【０１９５】

シンプルベイズ法において、文脈ｂで分類ａを出力する確率は、以下の数式１で与えられる。

【数１】

$$p(a|b)=\frac{p(a)}{p(b)}\,p(b|a)$$

$$\cong\frac{\tilde{p}(a)}{p(b)}\prod_i\tilde{p}(f_i|a)$$

【０１９６】

ただし、ここで文脈ｂは、あらかじめ設定しておいた素性f_j（∈Ｆ，１≦ｊ≦ｋ）の集合である。ｐ（ｂ）は、文脈ｂの出現確率である。ここで、分類ａに非依存であって定数のために計算しない。Ｐ（ａ）（ここでＰはｐの上部にチルダ）とＰ（f_i｜ａ）は、それぞれ教師データ（判断情報と同意義）から推定された確率であって、分類ａの出現確率、分類ａのときに素性f_iを持つ確率を意味する。Ｐ（f_i｜ａ）として最尤推定を行って求めた値を用いると、しばしば値がゼロとなり、数式２の２行目の式の値がゼロで分類先を決定することが困難な場合が生じる。そのため、スームージングを行う。ここでは、以下の数式２を用いてスームージングを行ったものを用いる。

【数２】

$$p(f_i|a)=\frac{freq(f_i,a)+0.01*freq(a)}{freq(a)+0.01*freq(a)}$$

【０１９７】

ただし、ｆｒｅｑ（f_i，ａ）は、素性f_iを持ちかつ分類がａである事例の個数、ｆｒｅｑ（ａ）は、分類がａである事例の個数を意味する。

【０１９８】

決定リスト法は、素性と分類先の組とを規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、検出する対象となる入力が与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性とを比較し、素性が一致した規則の分類先をその入力の分類先とする方法である。

【０１９９】

決定リスト方法では、あらかじめ設定しておいた素性f_j(∈Ｆ，１≦ｊ≦ｋ）のうち、いずれか一つの素性のみを文脈として各分類の確率値を求める。ある文脈ｂで分類ａを出力する確率は以下の数式３によって与えられる。

【数３】

$$p(a|b)=p(a|f_{max})$$

【０２００】
ただし、f_{max}は以下の数式４によって与えられる。
【数４】

$$f_{max}=\mathrm{argmax}_{fj\in F}\max_{ai\in A}\tilde{p}(a_i|f_j)$$

【０２０１】
また、Ｐ（a_i｜f_j）（ここでＰはｐの上部にチルダ）は、素性f_jを文脈に持つ場合の分類a_iの出現の割合である。
【０２０２】
最大エントロピー法は、あらかじめ設定しておいた素性f_j（１≦ｊ≦ｋ）の集合をＦとするとき、以下の所定の条件式（数式５）を満足しながらエントロピーを意味する式（数式６）を最大にするときの確率分布ｐ（ａ，ｂ）を求め、その確率分布にしたがって求まる各分類の確率のうち、最も大きい確率値を持つ分類を求める分類先とする方法である。
【数５】

$$\sum_{a\in A,b\in B}p(a,b)g_j(a,b)=\sum_{a\in A,b\in B}\tilde{p}(a,b)g_j(a,b)$$

$$\mathrm{for}\,\forall f_j(1\leq j\leq k)$$

【数６】

$$H(p)=-\sum_{a\in A,b\in B}p(a,b)\log(p(a,b))$$

【０２０３】
ただし、Ａ、Ｂは分類と文脈の集合を意味し、g_j（ａ，ｂ）は文脈ｂに素性f_jがあって、なおかつ分類がａの場合１となり、それ以外で０となる関数を意味する。また、Ｐ（a_i｜f_j）（ここでＰはｐの上部にチルダ）は、既知データでの（ａ，ｂ）の出現の割合を意味する。
【０２０４】
数式５は、確率ｐと出力と素性の組の出現を意味する関数ｇをかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として、エントロピー最大化(確率分布の平滑化)を行なって、出力と文脈の確率分布を求めるものとなっている。最大エントロピー法の詳細については、以下の参考文献（１）および参考文献（２）に記載されている。
【０２０５】
参考文献（１）：Eric Sven Ristad, Maximum Entropy Modeling for Natural Languag e,(ACL/EACL Tutorial Program, Madrid, 1997
【０２０６】
参考文献（２）：Eric Sven Ristad, Maximum Entropy Modeling Toolkit, Release1.6 beta, (http://www.mnemonic.com/software/memt,1998))
【０２０７】
サポートベクトルマシン法は、空間を超平面で分割することにより、二つの分類からなるデータを分類する手法である。
【０２０８】
図５にサポートベクトルマシン法のマージン最大化の概念を示す。図５において、白丸は正例、黒丸は負例を意味し、実線は空間を分割する超平面を意味し、破線はマージン領域の境界を表す面を意味する。図５（Ａ）は、正例と負例の間隔が狭い場合（スモールマージン）の概念図、図５（Ｂ）は、正例と負例の間隔が広い場合（ラージマージン）の概念図である。
【０２０９】

このとき、二つの分類が正例と負例からなるものとすると、学習データにおける正例と負例の間隔（マージン)が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ、図５（Ｂ）に示すように、このマージンを最大にする超平面を求めそれを用いて分類を行なう。

【０２１０】

基本的には上記のとおりであるが、通常、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線型にする拡張（カーネル関数の導入)がなされたものが用いられる。

【０２１１】

この拡張された方法は、以下の識別関数（ｆ（ｘ））を用いて分類することと等価であり、その識別関数の出力値が正か負かによって二つの分類を判別することができる。

【数７】

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_i K(x_i,x)+b\right)$$

$$b=-\frac{\max_{i,y_j=-1} b_i+\min_{i,y_j=1} b_i}{2}$$

$$b_i=\sum_{j=1}^{l}\alpha_j y_j K(x_j,x_i)$$

【０２１２】

ただし、ｘは識別したい事例の文脈（素性の集合)を、x_iとy_j（ｉ＝１，…，ｌ，ｙｊ∈｛１，－１｝）は学習データの文脈と分類先を意味し、関数ｓｇｎは、

ｓｇｎ（ｘ）＝１（ｘ≧０）

－１（otherwise）

であり、また、各α_iは数式８の式（８－２）と式（８－３）の制約のもと、式（８－１）を最大にする場合のものである。

【数８】

$$L(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j) \qquad \text{式（８－1）}$$

$$0\leq\alpha_i\leq C \quad (i=1,\cdots,l) \qquad \text{式（８－2）}$$

$$\sum_{j=1}^{l}\alpha_i y_i=0 \qquad \text{式（８－3）}$$

【０２１３】

また、関数Ｋはカーネル関数と呼ばれ、様々なものが用いられるが、本形態では、例えば、以下の多項式（数式９）のものを用いる。

【数９】

$$K(x,y)=(x\cdot y+1)d$$

【０２１４】

数式８、数式９において、Ｃ、ｄは実験的に設定される定数である。例えば、Ｃはすべての処理を通して１に固定した。また、ｄは、１と２の二種類を試している。ここで、$\alpha_i>0$となるx_iは、サポートベクトルと呼ばれ、通常、数式７の和をとっている部分は、この事例のみを用いて計算される。つまり、実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。

【０２１５】

なお、拡張されたサポートベクトルマシン法の詳細については、以下の参考文献（３）および参考文献（４）に記載されている。

【０２１６】

参考文献（３）：Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods,(Cambridge University Press,2000)

【０２１７】

参考文献（４）：Taku Kudoh, Tinysvm:Support Vector machines,(http://cl.aistnara.ac.jp/taku-ku//software/Tiny SVM/index.html,2000)

【０２１８】

サポートベクトルマシン法は、分類の数が２個のデータを扱うものである。したがって、分類の数が３個以上の事例を扱う場合には、通常、これにペアワイズ法またはワンＶＳレスト法などの手法を組み合わせて用いることになる。

【０２１９】

ペアワイズ法は、ｎ個の分類を持つデータの場合に、異なる二つの分類先のあらゆるペア（ｎ（ｎ－１）／２個）を生成し、各ペアごとにどちらがよいかを二値分類器、すなわちサポートベクトルマシン法処理モジュールで求めて、最終的に、ｎ（ｎ－１）／２個の二値分類による分類先の多数決によって、分類先を求める方法である。

【０２２０】

ワンＶＳレスト法は、例えば、ａ、ｂ、ｃという三つの分類先があるときは、分類先ａとその他、分類先ｂとその他、分類先ｃとその他、という三つの組を生成し、それぞれの組についてサポートベクトルマシン法で学習処理する。そして、学習結果による推定処理において、その三つの組のサポートベクトルマシンの学習結果を利用する。推定するべき問題が、その三つのサポートベクトルマシンではどのように推定されるかを見て、その三つのサポートベクトルマシンのうち、その他でないほうの分類先であって、かつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を求める解とする方法である。例えば、ある解くべき問題が、「分類先ａとその他」の組の学習処理で作成したサポートベクトルマシンにおいて分離平面から最も離れた場合には、その解くべき問題の分類先は、aと推定する。

【０２２１】

機械学習部１６が推定する、解くべき問題についての、どのような解（分類先）になりやすいかの度合いの求め方は、機械学習部１６が機械学習の手法として用いる様々な方法によって異なる。

【０２２２】

例えば、本発明の実施の形態において、機械学習部１６が、機械学習の手法としてｋ近傍法を用いる場合、機械学習部１６は、教師データの事例同士で、その事例から抽出された素性の集合のうち重複する素性の割合（同じ素性をいくつ持っているかの割合）にもとづく事例同士の類似度を定義して、前記定義した類似度と事例とを学習結果情報として学習データ格納部１２に記憶しておく。

【０２２３】

そして、機械学習部１６は、素性取得部１５によって解くべき問題の素性が抽出されたときに、記憶された類似度と事例を参照して、素性取得部１５によって抽出された解くべき問題の素性について、その解くべき問題の素性の類似度が高い順にｋ個の事例を選択し、選択したｋ個の事例での多数決によって決まった分類先を、解くべき問題の分類先（解）として推定する。すなわち、機械学習部１６では、解くべき問題についての、どのような解（分類先）になりやすいかの度合いを、選択したｋ個の事例での多数決の票数とする。

【０２２４】

また、機械学習手法として、シンプルベイズ法を用いる場合には、教師データの事例について、前記事例の解と素性の集合との組を学習データとして学習データ格納部１２に記

憶する。そして、機械学習部１６は、素性取得部１５によって解くべき問題の素性が抽出されたときに、学習データ格納部１２の判断情報の解と素性の集合との組をもとに、ベイズの定理にもとづいて素性取得部１５で取得した解くべき問題の素性の集合の場合の各分類になる確率を算出して、その確率の値が最も大きい分類を、その解くべき問題の素性の分類（解）と推定する。すなわち、機械学習部１６では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。

【０２２５】

また、機械学習手法として決定リスト法を用いる場合には、教師データの事例について、素性と分類先との規則を所定の優先順序で並べたリストを、予め、何らかの手段により、学習データ格納部１２に記憶させる。そして、素性取得部１５によって解くべき問題の素性が抽出されたときに、機械学習部１６は、学習データ格納部１２のリストの優先順位の高い順に、抽出された解くべき問題の素性と規則の素性とを比較し、素性が一致した規則の分類先をその解くべき問題の分類先（解）として推定する。

【０２２６】

また、機械学習手法として最大エントロピー法を使用する場合には、教師データの事例から解となりうる分類を特定し、所定の条件式を満足し、かつエントロピーを示す式を最大にするときの素性の集合と解となりうる分類の二項からなる確率分布を求めて、学習データ格納部１２に記憶する。そして、素性取得部１５によって解くべき問題の素性が抽出されたときに、機械学習部１６は、学習データ格納部１２の確率分布を利用して、抽出された解くべき問題の素性の集合についてその解となりうる分類の確率を求めて、最も大きい確率値を持つ解となりうる分類を特定し、その特定した分類をその解くべき問題の解と推定する。すなわち、機械学習部１６では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、各分類になる確率とする。

【０２２７】

また、機械学習手法としてサポートベクトルマシン法を使用する場合には、教師データの事例から解となりうる分類を特定し、分類を正例と負例に分割して、カーネル関数を用いた所定の実行関数にしたがって事例の素性の集合を次元とする空間上で、その事例の正例と負例の間隔を最大にし、かつ正例と負例を超平面で分割する超平面を求めて学習データ格納部１２に記憶する。そして、素性取得部１５によって解くべき問題の素性が抽出されたときに、機械学習部１６は、学習データ格納部１２の超平面を利用して、解くべき問題の素性の集合が超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果にもとづいて定まる分類を、その解くべき問題の解と推定する。すなわち、機械学習部１６では、解くべき問題の素性の集合の場合にある解となりやすさの度合いを、分離平面からのその解くべき問題の事例への距離の大きさとする。

（実験結果１）

【０２２８】

以下、異表記取得装置１の実験結果について説明する。まず、実験に利用するデータについて説明する。

【０２２９】

実験で用いるデータは、大規模類似語リストである。大規模類似語リストとは、検索エンジン研究基盤ＴＳＵＢＡＫＩ（http://tsubaki.ixnlp.nii.ac.jp/se/index.cgi参照［平成２１年１２月１３日検索］）の約１億ページ・６０億文のデータから１００万語を抽出し、その１００万語の各々の語に対して最大５００個の類義語を類似度付きで生成したものである。この大規模類似語リストに含まれる１００万語の日本語用語と、その日本語用語の各々の類義語の日本語用語を日本語用語対とする。

【０２３０】

そして、大規模類似語リストから、編集距離が１の日本語用語対をランダムに１４１８５組取り出した。その取り出した日本語用語対が日本語異表記対であるか、日本語異表記対でないかのタグ付けを３人の評定者の多数決により行った。３人の評定者のタグ付けがどれくらい一致しているのか、カッパ統計量Ｋを用いて判定する。

【０２３１】
右の用語と左の用語の２つの用語を有する日本語用語対の組み合わせが左にあるか、右にあるかにより、異なる情報になる素性がある。その素性に対応し、情報量を増やすために、日本語用語対の組み合わせを左右入れ替えたデータも用いる。つまり、本実験では、大規模類似リストから取り出した１４１８５組に加え、合計２８３７０組の実験データを用いる。
【０２３２】
また、２８３７０組の実験データを１つのまとまったデータであるとすると、実験の公正性が失われるのではないかと考え、２８４７０組ある実験データの半分を素性の考案を行うデータとして用いる。残り半分の実験データをクローズドデータで考案された素性が、他のデータにおいても有効であるかどうかの検討を行うデータとして用いる。素性の考案を行うデータをクローズドデータと呼ぶ。検討を行うデータをオープンデータと呼ぶ。クローズドデータは１０分割クロスバリデーションによる学習により精度の測定を行う。なお、１０分割クロスバリデーションとは、実験対象のデータを、第一から第十の１０に分割し、以下の（１）から（１０）の学習を行う。（１）第一をテストデータとし、第二から第十を学習データとして、学習を行う。（２）第二をテストデータとし、第一、第三から第十を学習データとして、学習を行う。（３）第三をテストデータとし、第一、第二、第四から第十を学習データとして、学習を行う。（４）第四をテストデータとし、他を学習データとして、学習を行う。（５）第五をテストデータとし、他を学習データとして、学習を行う。（６）第六をテストデータとし、他を学習データとして、学習を行う。（７）第七をテストデータとし、他を学習データとして、学習を行う。（８）第八をテストデータとし、他を学習データとして、学習を行う。（９）第九をテストデータとし、他を学習データとして、学習を行う。（１０）第十をテストデータとし、他を学習データとして、学習を行う。なお、１０分割クロスバリデーションは、公知技術である。
【０２３３】
また、クローズドデータを学習データ（学習データ格納部１２に格納されるデータ）、オープンデータをテストデータ（異表記の用語対であるか否かを判断されるデータ）とし、オープンクローズによる学習により精度の測定を行う。
【０２３４】
図６に、実験で用いた編集距離が１の日本語用語対の中に、多数決により日本語異表記対であるか日本語異表記対でないかを判定した内訳を示す。
【０２３５】
なお、カッパ統計量Ｋとは、Ｋ人評定者のカテゴリ評定における一致度を表す数値のことである。カッパ統計量Ｋの算出方法は公知であるので、説明を省略する。
【０２３６】
評定者の間に完全な一致があればＫの値は１になる。チャンスレベルでの一致であればＫの値は０である。一致度が高くなればＫの値は０から１に近づく。図７に、Ｌａｎｄｉｓらによる一致度の評価方法を示す。本実験では、１４１８５組の日本語用語対を対象に、３人で日本語異表記対であるか日本語異表記対でないかの２カテゴリでカッパ統計量Ｋを求めたところ、一致度は０．８４であった。これは０．８以上の一致度であるため、ほぼ完全な一致であると評価できる。
【０２３７】
次に、異表記取得装置１における異表記の用語対であるか否かの判断手法が優れていることを示すために、異表記取得装置１の判断手法と比較対照となるベースライン手法について説明する。
【０２３８】
編集距離の小さい（例えば、編集距離が１）日本語異表記対の抽出を行う対象の日本語用語対に対して、ベースライン手法では、以下のルールを適用し、機械的に日本語異表記対であるか日本語異表記対でないかについて判定を行う。
（ルール１）文字数が同じ日本語用語対の編集箇所が同じ値の数字を表す場合、日本語異

表記対であると判定する。
（ルール２）文字数が同じ日本語用語対の編集箇所が同じ意味のアルファベットを表す場合、日本語異表記対であると判定する。
（ルール３）文字数が同じ日本語用語対がＪＵＭＡＮを使い読み方を調べることでき、読み方が一致する場合、日本語異表記対であると判定する。
（ルール４）ルール１、ルール２、ルール３と一致しなかった場合、日本語異表記対でないと判定する。
【０２３９】
上記のルール１からルール４を適用するベースライン手法において、日本語用語対「第２版」「第二版」については、以下のように判断される。この日本語用語対における編集箇所は、「２」と「二」であり数字を表している。ルール１を適用し、同じ値を表しているため、ベースライン手法では、この日本語用語対は日本語異表記対であると判定される。
【０２４０】
日本語用語対「Ｔｅａ」「ｔｅａ」については、以下のように判断される。この日本語用語対における編集箇所は、「Ｔ」と「ｔ」でありアルファベットを表している。そして、ルール２が適用され、この日本語用語対は、同じ意味の語であるためこの日本語用語対は日本語異表記対であると判定される。
【０２４１】
日本語用語対「誉める」「褒める」については、以下のように判断される。この日本語用語対は、「ほめる」と「ほめる」にＪＵＭＡＮを使い読み方を調べることできる。そして、ルール３が適用され、読み方が一致し、この日本語用語対は日本語異表記対であると判定される。
【０２４２】
日本語用語対「シルキーホワイト」「ミルキーホワイト」については、以下のように判断される。この日本語用語対における編集箇所は、「シ」と「ミ」でありカタカナを表している。シルキーホワイトとミルキーホワイトはＪＵＭＡＮ辞書において未定義であるため、読み方を調べることができない。よってルール４が適用され、日本語異表記対でないと判定される。
【０２４３】
次に、本実験で用いた機械学習部１６の機械学習手法について、詳細に説明する。本機械学習手法は、サポートベクトルマシン法である。サポートベクトルマシン法は、上述したように、空間を超平面で分割することにより、２つの分類からなるデータを分類する手法である。このとき２つの分類が正例と負例からなるとすると、学習データにおいてこの２つの間隔が大きいものほど誤った分類をする可能性が低いと判断される。この間隔を最大にする超平面を求め、それを求めて分類を行うことが基本とされる。しかし、ここでは、学習データにおいて間隔の内部領域に少数の事例を含んでもよいとする手法や超平面の線形の部分を非線形にするなどの拡張がされたものを用いる。これらの拡張された方法は、識別関数を用いて分類することと等価となり、識別関数の出力値が正か負かによって２つに分類を判別することができる。また、３つ以上からなるデータを扱う場合にはペアワイズ手法というのを並行して用いる。ペアワイズ手法はＮ個の分類をもつデータの場合、異なる２つの分類先のあらゆるペアを作り、各ペアごとにどちらがよいかを２値分類器（サポートベクトルマシン法）で求め最終的に分類先の多数決により求める方法である。以降、サポートベクトルマシン法はＳＶＭと、適宜、表記する。
【０２４４】
ＳＶＭによって編集距離の小さい（例えば、編集距離が１）日本語異表記対を抽出するために用いる素性は、上述したＳ１からＳ６８の素性である。これらの素性は、大規模類似語リストからランダムで取り出した編集距離の小さい日本語用語対から取り出す。素性によってそれぞれの機械学習は、日本語用語対が日本語異表記対であるか日本語異表記対でないかを判定をする。日本語用語対からできるだけ多くの情報を得るために、種々の素

性を用いた。また、それぞれの素性について、上述したように、Ｇ１あらＧ７に分類できる。Ｇ１、Ｇ２、Ｇ３は、すべての編集距離の小さい日本語用語対に対応できる素性である。字種は対象の文字がひらがな、カタカナ、数字、アルファベット、その他のどの種類を表しているかの情報である。品詞は用語にＪＵＭＡＮを用いて形態素解析をかけ、用語を単語に区切り、品詞情報を取得する。そして、対象の文字がどの品詞に属しているかの情報である。位置情報は、対象の文字が品詞に属している中でさらに、その品詞の先頭、最後尾、それ以外のどの位置を示しているかの情報である。類似度は、大規模類似語リストを生成する際に用いた類似度の情報である。

【０２４５】

スタッキングアルゴリズムとは、上述したように、実験データを本来の目的とは別の分類方法で分類させたデータを機械学習で学習させ、学習結果の分類情報を素性に加えることである。本実験において、スタッキングアルゴリズムに使用するデータは、実験で用いる２８３７０組の日本語用語対以外の、大規模類似語リストから得られたＪＵＭＡＮの代表表記が判別できる９０４６１２組の日本語用語対を用いる。９０４６１２組の中で、正例は２５９３４組、負例は８７８６７８組である。これにより、ＪＵＭＡＮ辞書において未定義の日本語用語対にも、近似的ではあるがＳ６４の素性の情報を付与することができる。また、Ｇ４、Ｇ５は、特徴がある編集距離の小さい日本語用語対に特化した素性である。Ｇ４は置換によって等しい文字列で、Ｇ５は削除によって等しい文字列になる日本語用語対が対象である。Ｇ６、Ｇ７は、ＪＵＭＡＮ辞書、日本語ワードネット辞書、ＥＤＲ辞書を用いた素性である。ＪＵＭＡＮ辞書については、未定義とされている用語が出てくる日本語用語対に対して、素性の情報は付与しないこととする。

【０２４６】

上述したＳ１からＳ６８の素性がどれくらい有効であるのかを有意差の分析により検討する。有意の検討はブートストラップ法を用いて求める。ブートストラップ法とは分類手法によって二つに分類されたデータを用いる。分類された二つのデータをそれぞれ、データ数（例えば、問いの数は１４００）は変えずに重複を許しランダムに取り出す（例えば、取り出したデータ数は１４００）。取り出したデータでそれぞれのＦ値を求め、それぞれのＦ値を比較する。取り出しとＦ値の比較をする工程を１００００回繰り返す。Ｆ値とは以下の数式２で定義される。すなわち再現率と適合率の調和平均である。

【数１０】

$$F\text{値}=\frac{2*\text{再現率}*\text{適合率}}{\text{再現率}+\text{適合率}}$$

【０２４７】

工程を１００００回繰り返し比較した結果が、どちらかの手法のＦ値よりも、もう一方の手法のＦ値の方が高い回数が９５００回（９５％）以上の場合、有意水準５％によりＦ値が高い方の手法は有意であるといえる。どちらも９５００回（９５％）以上ない場合、有意水準５％により有意かどうかの判定はできない。なお、ここでは、有意水準５％を適用するが、例えば、有意水準１０％を適用しても良い。

【０２４８】

また、本実験では全素性と全素性から１種類の素性だけを取り除いたデータを、ＳＶＭの学習結果により比較する。上述した６８個すべて組み合わせた素性を全素性とし、取り除く素性はＳ１からＳ６８におけるすべての素性でおこなう。この有意差の検討を、クローズドデータを１０分割クロスバリデーション（１０ＣＶ）でＳＶＭによる学習結果と、クローズドデータとオープンデータを使いオープンクローズ（ＯＣ）でＳＶＭによる学習結果で行う。なお、オープンクローズとは、クローズドデータを学習データとして、オープンデータをテストデータとして実験することをいう。以降は、１０分割クロスバリデーションによるＳＶＭの実験は１０ＣＶと表記し、オープンクローズによるＳＶＭの実験はＯＣと表記する。

【０２４９】

次に、ベースライン手法と機械学習を利用した手法について実験を行った結果について報告する。

【０２５０】

本実験において、上述した大規模類似語リストに含まれる２８３７０組の編集距離の小さい日本語用語対が、日本語異表記対であるか、日本語異表記対でないかについて判定を行った。図８に、用意したクローズドデータとオープンデータに対して、ベースラインの手法を適用した結果を示す。また、図８には、１０ＣＶとＯＣの結果も示す。実験で用いるＳＶＭの実装としてＴｉｎｙＳＶＭを採用し、１次の多項式カーネルでソフトマージンパラメータＣを１に設定して利用した。それぞれの表での「全素性」はＳ１からＳ６８のすべての素性を利用した実験を示し、「素性選択」は省いた素性以外の全素性を利用した実験を示す。

【０２５１】

図８における正解率は、それぞれの実験データに対して、編集距離が１の日本語異表記対であるのか、編集距離が１の日本語異表記対でないのかを、正しく判定した割合である。図８のＦ値は、それぞれの実験データに対して、編集距離が１の日本語異表記対を抽出する場合のＦ値である。１０ＣＶ、ＯＣに対して、全素性を利用したＳＶＭの正解率、Ｆ値ともにベースラインの手法よりも高いことがわかる。編集距離１の日本語用語対から日本語異表記対を抽出する場合のＦ値は、ベースラインと比較して全素性を利用したＳＶＭの方が、１０ＣＶでは０．４３３高く、オープンデータでは０．４６０高かった。ベースラインの結果より、編集距離が１の日本語用語対から日本語異表記対を抽出することは難しいといえるが、本報告で提案している種々の素性と機械学習を用いた手法は、ベースライン手法よりも多くの日本語異表記対が抽出できることがわかる。

【０２５２】

次に、上述したブートストラップ法を用いて素性が有効であるかどうかの検討をした結果を図９に示す。なお、Ｓ１からＳ６８の素性を図１０に示す。図９において、省いた素性は、ブートストラップ法により有意かどうかの判定が行われる素性である。全素性は本実験で扱った素性による手法であり、素性選択は省いた素性を全素性から省いた素性による手法である。それぞれの値は、全素性が素性選択よりＦ値が高かった回数あるいは、素性選択が全素性よりＦ値が高かった回数である。この実験では、全素性が素性選択よりＦ値が高い回数が９５００回（９５％）以上あれば、省いた素性は精度向上に役立っているということになり、省いた素性は有効であるといえる。全素性が素性選択よりＦ値が高い回数が９５００回（９５％）以上であった素性は、１０ＣＶの場合はＳ４７、Ｓ５５、Ｓ５８、Ｓ６７であり、ＯＣの場合はＳ５２、Ｓ５４、Ｓ５５、Ｓ５８、Ｓ６７であった。１０ＣＶとＯＣの両方で、全素性が素性選択よりＦ値が高い回数が９５００回（９５％）以上であった素性は、Ｓ５５、Ｓ５８、Ｓ６７であった。この結果からＳ５５、Ｓ５８、Ｓ６７の素性は、どのような編集距離の小さい日本語異表記対を抽出するデータにも、有効である素性といえる。Ｓ４７とＳ５２の素性はそれぞれの実験で使われたデータには有効である素性といえるが、編集距離の小さい日本語用語対のデータが変われば、有効でなくなる可能性がある素性といえる。そのためＳ４７とＳ５２は、どのような編集距離の小さい日本語異表記対を抽出するデータにも、有効であるとはいえない。

【０２５３】

次に、本異表記取得装置１の提案手法（以下、単に提案手法とも言う）が、編集距離が１の日本語異表記対を抽出できたのかを、種々の同義語辞書を用いて比較を行った結果について説明する。種々の同義語辞書は、ＥＤＲ辞書、日本語ワードネット辞書、ＪＵＭＡＮ辞書である。編集距離が１の日本語用語対は、ＥＤＲ辞書には２１２２４７７９組、日本語ワードネット辞書には８９０６１６組、ＪＵＭＡＮ辞書には２３３４８組あることがわかった。ＥＤＲ辞書には人名に関する単語がある。本実験では人名は同義語でないと判断し、取り除いた。その結果、ＥＤＲ辞書に含まれている編集距離が１の日本語用語対は９３３０３７組であった。ＪＵＭＡＮ辞書は同じ代表表記をもつ単語対を日本語用語対として扱った。

【０２５４】
提案手法を用い、大規模類似語リストから編集距離が１の日本語異表記対と分類された用語対が、種々の辞書にどの程度の割合で含まれているかの検討結果を図１１に示す。以降は大規模類似語リストにおける編集距離が１の日本語用語対すべてを、日本語用語対データベースとし、用語対ＤＢと表記する。さらに、用語対ＤＢにおいて、日本語異表記対であると提案手法が分類した日本語用語対すべてを、日本語異表記対データベースとし、異表記ＤＢと表記する。用語対ＤＢにおいて、日本語異表記対でないと提案手法が分類した非日本語用語対すべてを、非日本語異表記対データベースとし、非異表記ＤＢと表記する。ＥＤＲ辞書は２０．４５％、日本語ワードネットは１．７１％、ＪＵＭＡＮ辞書は６．５２％の割合で異表記ＤＢの日本語異表記対が含まれていた。どの辞書においても、異表記ＤＢの日本語用語対を含んでいる割合は高くない。これらの結果より、本明細書で記載した異表記取得装置により得られた異表記と既存辞書は重なりが小さいので、異表記取得装置により、既存辞書に対して多くの異表記を追加できることが分かる。また、例えば、ＥＤＲ辞書では、約２割のカバー率であるが、相当な程度のカバー率である、と言える。

【０２５５】
また、種々の辞書に含まれる編集距離が１の日本語用語対を、提案手法により編集距離が１の日本語異表記対であるか、編集距離が１の日本語異表記対でないか分類した。ＳＶＭの分類における正解率を図１２に示す。

【０２５６】
また、種々の辞書と用語対ＤＢにおいて、編集距離が１の日本語異表記対であると分類された日本語用語対と、分類されなかった日本語用語対をそれぞれランダムに、５組ずつ取り出した結果を図１３に示す。学習データはオープンデータとクローズデータを組み合わせたデータとし、テストデータを用語対ＤＢ、種々の辞書のそれぞれでＯＣにより、用語対ＤＢと種々の辞書を分類した。

【０２５７】
図１２において、日本語ワードネットにおいて分類の正解率が低かったのは、図１３のように、日本語異表記対ではなく、日本語類義語対が多く含まれているからである。図１２に示すように、ＪＵＭＡＮ辞書の場合は８割という高い正解率で分類できている。また、ＪＵＭＡＮ辞書には、日本語異表記対でないものが含まれるという問題が少なく、また、本提案手法により適切に異表記を抽出できるために、８割という高い正解率を達成できたものと考えられる。

【０２５８】
次に、編集距離が１の日本語異表記対抽出の評価について述べる。ＳＶＭは識別関数の出力値（機械学習部１６が出力するスコア）が正か負かによって、データを分類することも可能であるが、ここでは、識別関数の出力値が正か負かによって、データを分類するのではなく、任意の値（閾値）によって正か負のデータを分類し、編集距離が１の日本語異表記対抽出の評価を行う。つまり、閾値判断手段１７２が、機械学習部１６が取得したスコアが閾値格納手段１７１に格納されている閾値以上または閾値より大きいか否かを判断するものとする。正のデータを編集距離が１の日本語異表記対であると分類し、負のデータを編集距離が１の日本語異表記対でないと分類する手法では、精度が１００％ではないため、誤って編集距離が１の日本語異表記対でない日本語用語対を、日本語異表記対であると判断し、抽出することがある。そのため、少量であっても確実に抽出を行いたい場合は、閾値を高く設定することで、日本語異表記対を確実に抽出できる。また、誤ったデータが含まれていても、網羅的に抽出を行いたい場合は、閾値を低く設定することで、可能となる。図１４に、閾値の評価基準を示す。図１４に示すように、閾値を－０．２に設定することで、Ｆ値０．９３２３と最も高い値を得られることがわかった。また、再現率と適合率の比率を図１５に示す。図１５によれば、再現率が高くしようとすると、カバー率を上げなくてはならなくなり、適合率は低くなる。そして、適合率が高くしようとすると、カバー率を下げなくてはならなくなり、再現率は低くなる。

（実験結果２）
【０２５９】
第２番目の実験において、正例（例えば、「スパゲティ」と「スパゲッティ」との対）７４５個、負例（正例に該当しない対）１３，４４０個を持つ学習データから、正例７２５個、負例１３，４６０個のテストデータの抽出が行なわれ、そのＦ値は、０．９３であった。なお、実験結果２において、実験結果１で利用した素性や学習データが完全に一致するものではないが、本提案手法の有効性を示すために足りる、素性や学習データの重複がある。
【０２６０】
すべてを正例と判断する、即ちどんなものでも正例とするベースラインの方法であると、Ｆ値は０．０９７２程度であった。異表記かどうかを判定する対象の用語対において、編集箇所の文字また、編集箇所の文字の周辺の文字だけの素性を用いる従来の方法でも、Ｆ値は０．８５であった。つまり、提案手法のように、多数の素性（ここでは、６８）を用いた方法の効果は顕著であることが分かる。
【０２６１】
また、既存の異表記辞書に基づく素性、また、スタッキング手法に基づく素性（上記の辞書関連素性）を利用しなかった方法よりも、これらの方法を利用した方が有意にＦ値が高いことも確かめており、これらの手法の有効性も確認している。
【０２６２】
また、ルールベース的手法として、編集箇所の文字の字種が漢数字かアラビア数字であること、または、同じアルファベットであること、また、既存の異表記辞書を利用することで異表記と判定できるものを、異表記と決定的に推定する方法も試した。この場合のＦ値は、０．４２０２であり、ルールベース的手法でなく教師あり機械学習を利用する方が良いことがわかる。
【０２６３】
正しい異表記の対の差分データから、異表記になりやすい差分パターンを学習し、ある用語Ａに対して、異表記の候補Ｂを上記差分パターンより生成し、用語Ａと用語Ｂが異表記の対であるかを判定する操作を利用することにより、取得できる異表記が格段に増えるという効果がある。かかる操作については、実施の形態２で説明する。
（実験結果３）
【０２６４】
第３番目の実験において、１０万語の単語とそれの類似する１００語の単語を用いた。１０万語の単語とそれの類似する１００語の単語のすべての対のうち、１文字のみ変化している用語対は１７０万個あった。なお、実験結果３において、実験結果１で利用した素性や学習データが完全に一致するものではないが、本提案手法の有効性を示すために足りる、素性や学習データの重複がある。そして、異表記取得装置１の技術を利用して、そこから７万対の異表記を取り出せる。以下に構築できる異表記の例を示す。
? ＢｕｓｉｎｅｓｓＷｅｅｋ　　Ｂｕｓｉｎｅｓｓ　Ｗｅｅｋ
? ＪＡＶＡＳｃｒｉｐｔ　　ＪＡＶＡ　Ｓｃｒｉｐｔ
? 書いてた頃　　書いていた頃
? アイリッシュトラッド　　アイリッシュ・トラッド
? 自サーバ　　自サーバー
? でない場合　　出ない場合
? ＷＷＷサーバ上　　ＷＷＷサーバー上
? 日光彫　　日光彫り
? 隣同士　　隣り同士
【０２６５】
なお、ＥＤＲ（Ｅｌｅｃｔｒｉｃ　Ｄｉｃｔｉｏｎａｒｙ　Ｒｅｓｅａｒｃｈ）電子化辞書に含まれる差分が１文字の異表記は２４，１８５語である。また、日本語ＷｏｒｄＮｅｔに含まれる差分が１文字の異表記のようなものは８２，２７０語ある。ただし、日本

語ＷｏｒｄＮｅｔには、異表記でないもの（類義語）も多く含まれており、適切に異表記を取り出すことが困難である。さらに、ＪＵＭＡＮの辞書に含まれる差分が１文字の異表記は２３，３４８語である。これらと比較しても本提案手法の技術の有効性がわかる。また、「ＪＡＶＡ」は登録商標です。

（実験結果４）

【０２６６】

第４番目の実験において、上記したベースライン手法（上記のルール１からルール４を適用した方法）による精度を算出する。ベースライン手法では、有意差が高かった素性（Ｓ５５、Ｓ５８、Ｓ６７）がｙｅｓと判定されたものを正例、すべてｎｏと判定されたものを負例としてＦ値を求める。

【０２６７】

図１６は、ベースライン手法で、オープンデータとクローズドデータの全部を用いて１０分割クロスバリデーションによる実験をおこなった場合の結果である。図１６において、「０」は負例、「１」は正例である。また、図１６において、最も左側の列の「０」「１」は、正しい分類を示す。最も上の第一行の「０」「１」は、実験対象の手法（図１６では、ベースライン手法）での出力結果を示す。つまり、正しい分類が「０」であり実験結果が「０」であったデータの数が２６８９２、正しい分類が「０」であり実験結果が「１」であったデータの数が１０１８、正しい分類が「１」であり実験結果が「０」であったデータの数が８、正しい分類が「１」であり実験結果が「１」であったデータの数が４５２であった。また、負例（「０」）の再現率は９９．９７％、適合率は９６．３５％であった。また、正例（「１」）の再現率は３０．７５％、適合率は９８．２６％であった。また、すべてのデータの再現率は９６．３８％、適合率は９６．３８％であった。さらに、「総数」は、実験データの数である。以上の再現率、適合率を、数式１０に代入して、算出した負例のＦ値は０．９８１３、正例のＦ値は０．４６８４であった。なお、図１７から図２１の各データの意味は、図１６と同様であるので説明を省略する。

【０２６８】

図１７は、ベースライン手法で、クローズドデータのみを用いて１０分割クロスバリデーションによる実験をおこなった場合の結果である。図１７において、負例のＦ値は０．９８１４、正例のＦ値は０．４８３３であった。

【０２６９】

図１８は、ベースライン手法で、オープンクローズを用いた場合の結果である。図１８において、負例のＦ値は０．９８１２、正例のＦ値は０．４５２９であった。

【０２７０】

実験結果４において、ベースライン手法は、正例のＦ値が、提案手法におけるＦ値（例えば、実験結果１の０．９１２）と比較して極めて小さく、提案手法の有効性が極めて高い、と言える。

（実験結果５）

【０２７１】

第５番目の実験において、すべてを正例としたベースライン手法の場合による精度を算出した。すべて正例としたベースライン手法の場合、再現率は「１００％」、適合率は「０．０５２５％」であった。そして、かかる再現率および適合率を、数式１０に代入し、算出された、正例（「１」）のＦ値は「０．０９９８」であった。すべてを正例としたベースライン手法の正例のＦ値は、提案手法におけるＦ値と比較して極めて小さく、提案手法の有効性が極めて高い、と言える。なお、提案手法において、正解率「９９．１２％」、再現率「９９．０７％」、適合率「９２．２９％」、Ｆ値「０．９１２」を得ている。なお、本実験で利用した素性や学習データは、提案手法の評価において利用した素性や学習データと完全に一致するものではないが、本提案手法の有効性を示すために足りる、素性や学習データの重複がある。

（応用例）

【０２７２】

以下、異表記取得装置１の応用例について説明する。応用例とは、異表記取得装置１を組み込んだ情報検索装置である。情報検索装置は、異表記取得装置１と検索部とを具備する。つまり、受付部１４は、キーワード（ＫＷ１）を受け付ける。そして、異表記取得装置１は、受け付けたＫＷ１の異表記の用語（ＫＷ２）を取得する。そして、検索部は、ＫＷ１＋ＫＷ２（＋はＯＲ）の検索式により、情報検索を行う。なお、情報検索の検索対象は問わないことは言うまでもない。また、検索部は、いわゆるＷｅｂの検索エンジンを起動するだけの処理でも良い。

【０２７３】

本情報検索装置を利用して、情報をキーワード検索する際に、ユーザが、「スパゲティ」と入力した場合に、情報検索装置は、「スパゲティ」の異表記である「スパゲッティ」を取得する。そして、情報検索装置は、これらの「スパゲティ」と「スパゲッティ」との双方をキーワードとして情報検索する。その結果、「スパゲティ」と「スパゲッティ」のいずれの表現が為されている情報もヒットするので、検索漏れの少ない情報検索が実現できる。

【０２７４】

特に情報検索装置は、検索漏れが許されない特許情報の検索に大きな効果をもたらす。例えば、情報検索装置が特許検索システムにおいて利用されることを考える。特許の明細書や特許請求の範囲や要約書等の特許の書類には、例えば、「コンピュータ」も「コンピューター」も存在するので、キーワードとして「コンピュータ＋コンピューター」を入力しなければ、検索漏れが生じる。従って、検索者は検索時には細心の注意を払って検索しようとするキーワードの異表記を考える必要があった。「デジタル」と「ディジタル」など、同義語であるにも拘わらず、異表記の文言は特に特許公報においては多い。しかしながら、本情報検索装置を採用することによってこのような配慮をすることなく、検索漏れのない特許情報の検索が可能となる。

【０２７５】

以上、本実施の形態によれば、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる。

【０２７６】

なお、本実施の形態によれば、主として、編集距離が１の用語対について、異表記の用語対であるか否かの判断手法について説明した。しかし、上述したとおり、異表記取得装置１は、編集距離が２の用語対についても、異表記の用語対であるか否かを判断できる。

【０２７７】

つまり、素性取得部１５の差分文字取得手段１５１は、編集距離が２つの用語対について、２つの差分文字の組を、それぞれ取得する。例えば、以下の３つの具体的な用語対を考える。（１）「できる」「出来る」（２）「理解できる」「できる」（３）「ＩＸ（ローマ数字の９）」「９」を考える。かかる場合、差分文字取得手段１５１は、（１）の用語対について、「で」「出」と「き」「来」の２組の差分文字の組を取得する。また、差分文字取得手段１５１は、（２）の用語対について、「理」「」と「解」「」（「」はＮＵＬＬである）の２組の差分文字の組を取得する。また、差分文字取得手段１５１は、（３）の用語対について、「Ｉ」「９」と「Ｘ」「」の２組の差分文字の組、または「Ｉ」「」と「Ｘ」「９」の２組の差分文字の組を取得する。

【０２７８】

そして、素性取得手段１５２は、差分文字取得手段１５１が取得した２つの差分文字を、独立に対象として、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を、２組取得する。つまり、（１）の用語対について、素性取得手段１５２は、「で」「出」と「き」「来」の２組の差分文字の組のそれぞれを対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、２種類のテストデータを作成する。素性取得手段１５２は、例えば、用語対が有する２つの用語の編集箇所の字種が異なり、かつ、２つの用語の編集箇所が同じ値の数字であるか否かを示す字種関連素性について、「で」「出」の編集箇所が同じ値の数字でないと判断し、当該字種関連素性「０」

を取得する。また、素性取得手段１５２は、例えば、「で」「出」について、２つの用語の読みが一致するか否かを示す辞書関連素性「１」を取得する。素性取得手段１５２は、用語辞書１３から「出」の読み「で」を取得し、「で」と「出」の読みが一致すると判断する。また、素性取得手段１５２は、例えば、差分文字「で」「出」に対して、差分文字（編集箇所）の前後の文字という素性について、前の文字の素性「」（なし）、後の文字の素性「き」と「来」を取得する。また素性取得手段１５２は、例えば、差分文字「き」「来」に対して、差分文字の前後の文字という素性について、前の文字の素性「出」と「で」、後の文字の素性「る」を取得する。かかる処理により、別の差分文字も素性に含めることとなる。

【０２７９】

また、（２）の用語対について、素性取得手段１５２は、（１）と同様に、「理」「」と「解」「」の２組の差分文字の組のそれぞれを対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、２種類のテストデータを作成する。さらに、（３）の用語対について、素性取得手段１５２は、（１）（２）と同様に、例えば、「Ｉ」「９」と「Ｘ」「」の２組の差分文字の組のそれぞれを対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、２種類のテストデータを作成する。

【０２８０】

次に、機械学習部１６は、（１）（２）（３）について、２種類のテストデータをそれぞれ、異表記の用語対であるか否かを判定する。そして、判定の結果、例えば、２種類のテストデータともに異表記の用語対であると判定された場合、元の用語対（例えば、「できる」「出来る」）は、異表記の用語対であるとして、出力部１７は、判断結果を出力する。なお、出力部１７は、上述したように、２種類のテストデータに対する２つのスコアのうちのスコアが０に近い方のスコアを採用して、採用したスコアが正の場合は正例（異表記の用語）、負の場合は負例（異表記の用語でない）と判断しても良いし、スコアの絶対値が大きい方のスコアを採用して、採用したスコアが正の場合は正例（異表記の用語）、負の場合は負例（異表記の用語でない）と判断しても良いし、２つのスコアのうち、小さい方のスコアを取得し、当該小さい方のスコアが正の場合は正例（異表記の用語）、負の場合は負例（異表記の用語でない）と判断しても良い。

【０２８１】

また、（３）の２組の差分文字の組（例えば、「Ｉ」「９」と「Ｘ」「」、または「Ｉ」「」と「Ｘ」「９」）、つまり２つの問題（問題１、問題２）ができる場合、それぞれの差分文字を対象に素性の抽出を行い、それぞれ差分文字から抽出した素性は、別のものと考え、４種類のテストデータを作成する。そして、２つの問題ごとに、算出したスコアが０に近い方を取得し、問題ごとのスコアのうちの、絶対値が高いスコアを当該問題のスコアとし、スコアが正の場合は正例、負の場合は負例と判断しても良い。また、例えば、用語対が（３）「ＩＸ」「９」である場合、問題「Ｉ」「９」と「Ｘ」「」、および「Ｉ」「」と「Ｘ」ができる。そして、機械学習部１６は、「Ｉ」「９」と「Ｘ」「」のスコアの小さい方を取得し、また、「Ｉ」「」と「Ｘ」「９」のスコアの小さい方を取得し、２つの取得されたスコアのうち、値が大きい方を「ＩＸ」「９」の用語対におけるスコアとする。そして、機械学習部１６は、当該スコアが正の場合は正例、負の場合は負例と判断しても良い。なお、例えば、機械学習部１６は、「Ｉ」「９」と「Ｘ」「」のスコアが０に近い方を取得し、また、「Ｉ」「」と「Ｘ」「９」のスコアが０に近い方を取得し、２つの取得されたスコアのうち、絶対値が大きい方を「ＩＸ」「９」の用語対におけるスコアとしても良い。そして、機械学習部１６は、当該スコアが正の場合は正例、負の場合は負例と判断しても良い。

【０２８２】

また、本実施の形態において、編集距離が３以上の用語対についても、編集距離が２つの用語対と同様に、３以上のテストデータを作成し、３以上のテストデータの判断結果を用いて、元の用語対が異表記の用語対であるか否かを判定しても良い。かかる場合、例えば、３以上の差分文字のうちの１文字や２文字などを素性として用いるなど、新しい素性

を機械学習手法に導入しても良い。

【０２８３】

また、本実施の形態において、異表記取得装置１は、例えば、「あなた」「あんた」という日本語の用語対が異表記の用語対であると判断できたが、日本語以外の言語（例えば、英語）の用語対（例えば、「colour」「color」）も、異表記の用語対であると判断できる。

【０２８４】

また、本実施の形態において、異表記取得装置１は、用語対を構成する２つの用語の編集箇所の文字が２文字以上である場合、１文字ずつの対応とせずに、編集箇所をまとめ、当該まとめた文字列をそのまま機械学習しても良い。つまり、用語対「１２３組」「百二十三組」について、編集箇所を「１２３」「百二十三」とまとめて、処理しても良い。用語対「１２３組」「百二十三組」に対して、例えば、Ｓ１「一つ目の表記の編集箇所」"１２３"、Ｓ２「二つ目の表記の編集箇所」"百二十三"、Ｓ３「編集箇所の前方の１文字」""（なし）、Ｓ４「編集箇所の後方の１文字」"組"、Ｓ５５「編集箇所が両方とも数字の場合であり、同じ値か違う値かどうか」"１"（同じ値）、Ｓ５６「日本語用語対の編集箇所が両方ともひらがなの場合であり、同じ音声か違う音声かどうか」"０"等が得られる。そして、用語対「１２３組」「百二十三組」に対する学習データが構成され、学習データ格納部１２に蓄積されて、利用されても良い。また、異表記取得装置１の素性取得部１５は、編集箇所の文字が２文字以上の用語対のテストデータに対して、編集箇所をまとめて処理し、例えば、上述した６８の素性を取得し、機械学習部１６が、テストデータが異表記対か否かを判断しても良い。

【０２８５】

さらに、本実施の形態における処理は、ソフトウェアで実現しても良い。そして、このソフトウェアをソフトウェアダウンロード等により配布しても良い。また、このソフトウェアをＣＤ－ＲＯＭなどの記録媒体に記録して流布しても良い。なお、このことは、本明細書における他の実施の形態においても該当する。なお、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、記憶媒体に、編集距離が１以上の用語対、および、用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、前記用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、前記用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納しており、コンピュータを、前記記憶媒体の用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上を含む複数の素性を取得する素性取得部と、前記各用語対に対して、前記記憶媒体の２以上の学習データと、前記素性取得部が取得した複数の素性とを用いて、教師あり機械学習法により、前記記憶媒体の各用語対が異表記の用語対であるか否かを判断する機械学習部と、前記機械学習部における判断結果を出力する出力部として機能させるためのプログラム、である。

（実施の形態２）

【０２８６】

本実施の形態において、置き換え文字対を保持し、置き換え文字対を用いて、用語から用語対を生成し、その用語対に対して、機械学習により異表記用語対を生成する異表記取得装置２について説明する。異表記取得装置２は、異表記取得装置１の機能に加えて、パターンを使った異表記用語対の生成機能を有する機能を有する。

【０２８７】

図１９は、本実施の形態における異表記取得装置２のブロック図である。

異表記取得装置２は、用語対格納部１１、異表記用語対格納部２１、学習データ格納部１２、用語辞書１３、異表記パターン格納部２２、受付部２３、編集箇所取得部２４、異表記パターン取得部２５、異表記パターン蓄積部２６、用語対生成部２７、素性取得部１５、機械学習部１６、出力部１７を備える。

【０２８８】
異表記用語対格納部２１は、編集距離が１の異表記の用語対を１以上格納し得る。
【０２８９】
異表記パターン格納部２２は、異表記のパターンを示す第一文字列と第二文字列とを対に有する異表記パターンを１以上格納し得る。
【０２９０】
受付部２３は、ユーザからの入力を受け付ける。また、受付部２３は、１以上の用語を受け付ける。この用語とは、用語対を生成する元となる用語である。受付部２３が用語を受け付けるのは、ユーザからの入力でも良いし、記憶媒体からの読み込みや、通信手段を用いた受信でも良い。
【０２９１】
編集箇所取得部２４は、異表記用語対格納部２１に格納されている１以上の異表記の用語対の編集箇所を取得する。
【０２９２】
異表記パターン取得部２５は、編集箇所取得部２４が取得した編集箇所から、第一文字列と第二文字列とを対に有する異表記パターンを取得する。異表記パターン取得部２５は、例えば、用語対「２番目」「二番目」から第一文字列「２」と第二文字列「二」とを対に有する異表記パターン「２」「二」を取得する。また、異表記パターン取得部２５は、例えば、用語対「自サーバ」「自サーバー」から異表記パターン「del」「ー」を取得する。
【０２９３】
異表記パターン蓄積部２６は、異表記パターン取得部２５が取得した異表記パターンを、異表記パターン格納部２２に蓄積する。
【０２９４】
用語対生成部２７は、受付部２３が受け付けた１以上の各用語に対して、異表記パターン格納部２２の１以上の各異表記パターンを適用し、１以上の用語を生成し、１以上の各用語と生成した用語とを有する１以上の異表記の候補の用語対である異表記候補用語対を生成する。
【０２９５】
異表記用語対格納部２１、異表記パターン格納部２２は、不揮発性の記録媒体が好適であるが、揮発性の記録媒体でも実現可能である。
【０２９６】
異表記用語対格納部２１に異表記用語対が記憶される過程は問わない。
【０２９７】
編集箇所取得部２４、異表記パターン取得部２５、異表記パターン蓄積部２６、および用語対生成部２７は、通常、ＭＰＵやメモリ等から実現され得る。編集箇所取得部２４等の処理手順は、通常、ソフトウェアで実現され、当該ソフトウェアはＲＯＭ等の記録媒体に記録されている。但し、ハードウェア（専用回路）で実現しても良い。
【０２９８】
次に、異表記取得装置２の動作について図２０のフローチャートを用いて説明する。図２０のフローチャートにおいて、異表記パターンを蓄積する処理、用語対を生成する処理について説明する。異表記取得装置２の動作について、異表記取得装置１の動作と同じである異表記用語対の判断処理、および判断結果の出力処理については、説明しない。
【０２９９】
（ステップＳ２００１）受付部２３は、異表記パターンの生成指示を受け付けたか否かを判断する。異表記パターンの生成指示を受け付ければステップＳ２００２に行き、受け付けなければステップＳ２００９に行く。
【０３００】
（ステップＳ２００２）編集箇所取得部２４は、カウンタｉに１を代入する。
【０３０１】

（ステップＳ２００３）編集箇所取得部２４は、異表記パターン格納部２２にｉ番目の用語対が存在するか否かを判断する。ｉ番目の用語対が存在すればステップＳ２００４に行き、ｉ番目の用語対が存在しなければステップＳ２００１に戻る。

【０３０２】

（ステップＳ２００４）編集箇所取得部２４は、ｉ番目の用語対の差分文字（編集箇所）を取得する。

【０３０３】

（ステップＳ２００５）異表記パターン取得部２５は、ステップＳ２００４で取得した差分文字（編集箇所）から、異表記パターンを構成する。

【０３０４】

（ステップＳ２００６）異表記パターン蓄積部２６は、ステップＳ２００５で取得された異表記パターンが、異表記パターン格納部２２に存在するか否かを判断する。存在すればステップＳ２００７に行き、存在しなければステップＳ２００８に行く。

【０３０５】

（ステップＳ２００７）異表記パターン蓄積部２６は、ステップＳ２００５で取得された異表記パターンを、異表記パターン格納部２２に蓄積する。

【０３０６】

（ステップＳ２００８）編集箇所取得部２４は、カウンタｉを１、インクリメントする。ステップＳ２００３に戻る。

【０３０７】

（ステップＳ２００９）受付部２３は、用語を受け付けたか否かを判断する。用語を受け付ければステップＳ２０１０に行き、受け付けなければステップＳ２００１に戻る。

【０３０８】

（ステップＳ２０１０）用語対生成部２７は、カウンタｉに１を代入する。

【０３０９】

（ステップＳ２０１１）用語対生成部２７は、ｉ番目の異表記パターンが、異表記パターン格納部２２に存在するか否かを判断する。存在すればステップＳ２０１２に行き、存在しなければ処理を終了する。

【０３１０】

（ステップＳ２０１２）用語対生成部２７は、ステップＳ２００９で受け付けられた用語が、ｉ番目の異表記パターンに合致するか否かを判断する。合致すればステップＳ２０１３に行き、合致しなければステップＳ２０１６に行く。なお、用語「ＷＷＷサーバ」に対して、異表記パターン「２」「二」は合致しない。異表記パターンは両方とも文字であり、当該いずれの文字も用語「ＷＷＷサーバ」が含まないからである。また、用語「ＷＷＷサーバ」に対して、異表記パターン「del」「ー」は合致する。異表記パターンに「del」が含まれる場合は、すべての用語が異表記パターンに合致することとなる。

【０３１１】

（ステップＳ２０１３）用語対生成部２７は、ステップＳ２００９で受け付けられた用語に対して、ｉ番目の異表記パターンを適用し、１以上の異表記の用語を取得する。用語が「アイトラッキング」であり、ｉ番目の異表記パターンが「del」「・」である場合、用語対生成部２７は、用語「アイトラッキング」に異表記パターン「del」「・」を適用し、「・」を各文字間に挿入し、７つの異表記の用語「ア・イトラッキング」「アイ・トラッキング」「アイト・ラッキング」「アイトラ・ッキング」「アイトラッ・キング」「アイトラッキ・ング」「アイトラッキン・グ」を生成する。また、用語が「一番目」であり、ｉ番目の異表記パターンが「一」「１」である場合、用語対生成部２７は、用語「一番目」に異表記パターン「一」「１」を適用し、「１番目」を生成する。

【０３１２】

（ステップＳ２０１４）用語対生成部２７は、ステップＳ２００９で受け付けられた用語と、ステップＳ２０１３で生成した１以上の異表記の用語を用いて、１以上の用語対を生成する。例えば、用語が「アイトラッキング」であり、ｉ番目の異表記パターンが「de

l」「・」である場合、用語対生成部２７は、用語対「アイトラッキング」「ア・イトラッキング」、「アイトラッキング」「アイ・トラッキング」、「アイトラッキング」、「アイトラッキング」「アイト・ラッキング」、「アイトラッキング」「アイトラ・ッキング」、「アイトラッキング」「アイトラッ・キング」、「アイトラッキング」「アイトラッキ・ング」、「アイトラッキング」「アイトラッキン・グ」の７つの用語対を生成する。また、用語が「一番目」であり、ｉ番目の異表記パターンが「一」「１」である場合、用語対生成部２７は、用語対「一番目」「１番目」を生成する。

【０３１３】

（ステップＳ２０１５）用語対生成部２７は、ステップＳ２０１３で生成した１以上の用語対を、用語対格納部１１に蓄積する。

【０３１４】

（ステップＳ２０１６）用語対生成部２７は、カウンタｉを１、インクリメントする。ステップＳ２０１１に戻る。

【０３１５】

以下、本実施の形態における異表記取得装置２の具体的な動作について説明する。

【０３１６】

異表記パターン取得部２５が取得し、異表記パターン蓄積部２６が異表記パターン格納部２２に蓄積した異表記パターンの例を、図２１に示す。図２１において、「ｄｅｌ」は、もう一方のパターン文字を削除することを示す。つまり、異表記パターン取得部２５は、delのもう一方のパターン文字について、すべての大規模類似語リストに用いた用語を対象とし、用語対を生成する。

【０３１７】

かかる状況において、上述したように、用語対生成部２７は、受け付けられた用語「アイトラッキング」が、１番目の異表記パターン「del」「・」に合致する、と判断する。そして、用語が「アイトラッキング」が入力された場合、１番目の異表記パターン「del」「・」が適用され、用語対生成部２７は、用語「アイトラッキング」は、「・」を各文字間に挿入し、７つの異表記の用語「ア・イトラッキング」「アイ・トラッキング」「アイト・ラッキング」「アイトラ・ッキング」「アイトラッ・キング」「アイトラッキ・ング」「アイトラッキン・グ」を生成する。次に、用語対生成部２７は、用語対「アイトラッキング」「ア・イトラッキング」、「アイトラッキング」「アイ・トラッキング」、「アイトラッキング」、「アイトラッキング」「アイト・ラッキング」、「アイトラッキング」「アイトラ・ッキング」、「アイトラッキング」「アイトラッ・キング」、「アイトラッキング」「アイトラッキ・ング」、「アイトラッキング」「アイトラッキン・グ」の７つの用語対を生成する。そして、異表記パターン蓄積部２６は、７つの用語対を異表記パターン格納部２２に蓄積する。

【０３１８】

次に、用語対生成部２７は、受け付けられた用語「アイトラッキング」が、２番目の異表記パターン「del」「－」に合致する、と判断する。次に、用語「アイトラッキング」に対して、２番目の異表記パターン「del」「－」が適用され、用語対生成部２７は、「ア－イトラッキング」「アイ－トラッキング」「アイト－ラッキング」「アイトラ－ッキング」「アイトラッ－キング」「アイトラッキ－ング」「アイトラッキン－グ」を生成する。次に、用語対生成部２７は、用語対「アイトラッキング」「ア－イトラッキング」、「アイトラッキング」「アイ－トラッキング」、「アイトラッキング」「アイト－ラッキング」、「アイトラッキング」「アイトラ－ッキング」、「アイトラッキング」「アイトラッ－キング」、「アイトラッキング」「アイトラッキ－ング」、「アイトラッキング」「アイトラッキン－グ」の７つの用語対を生成する。そして、異表記パターン蓄積部２６は、７つの用語対を異表記パターン格納部２２に蓄積する。

【０３１９】

次に、用語対生成部２７は、受け付けられた用語「アイトラッキング」が、３番目の異表記パターン「del」「い」に合致する、と判断する。そして、次に、用語「アイトラッ

キング」に対して、３番目の異表記パターン「del」「い」が適用され、用語対生成部２７は、「アいイトラッキング」「アイいトラッキング」「アイトいラッキング」「アイトラいッキング」「アイトラッいキング」「アイトラッキいング」「アイトラッキンいグ」を生成する。次に、用語対生成部２７は、用語対「アイトラッキング」「アいイトラッキング」、「アイトラッキング」「アイいトラッキング」、「アイトラッキング」「アイトいラッキング」、「アイトラッキング」「アイトラいッキング」、「アイトラッキング」「アイトラッいキング」、「アイトラッキング」「アイトラッキいング」、「アイトラッキング」「アイトラッキンいグ」の７つの用語対を生成する。そして、異表記パターン蓄積部２６は、７つの用語対を異表記パターン格納部２２に蓄積する。

【０３２０】

次に、用語対生成部２７は、受け付けられた用語「アイトラッキング」が、４番目の異表記パターン「ー」「１」を構成する文字を含まないので、この異表記パターンに合致しない、と判断する。

【０３２１】

次に、用語対生成部２７は、受け付けられた用語「アイトラッキング」が、５番目の異表記パターン「イ」「ィ」を構成する文字「イ」を含むので、この異表記パターンに合致する、と判断する。そして、次に、用語「アイトラッキング」に対して、５番目の異表記パターン「イ」「ィ」が適用され、用語対生成部２７は、「アィトラッキング」を生成する。次に、用語対生成部２７は、用語対「アイトラッキング」「アィトラッキング」の１つの用語対を生成する。そして、異表記パターン蓄積部２６は、１つの用語対を異表記パターン格納部２２に蓄積する。

【０３２２】

次に、同様に、用語対生成部２７は、６番目以降の異表記パターンを適用して、処理していく。

【０３２３】

そして、用語対生成部２７は、新たな用語対を用語対格納部１１に蓄積する。

【０３２４】

以上、本実施の形態によれば、異表記の用語対の候補を自動生成できる。また、本実施の形態によれば、異表記の用語対の候補を自動生成するための異表記パターンを自動的に取得できる。

【０３２５】

なお、本実施の形態における情報処理装置を実現するソフトウェアは、以下のようなプログラムである。つまり、このプログラムは、記憶媒体に、用語対の異なる文字である編集箇所の字種に関する素性である字種関連素性、用語辞書を用いて取得された素性である辞書関連素性、前記用語対を構成する２つの用語の類似度を示す素性である類似度素性のうちの一以上の素性を含む複数の素性と、前記用語対が異表記の用語対であるかを示す情報である正負情報とを対応付けた学習データを２以上格納しており、記憶媒体に、異表記のパターンを示す第一文字列と第二文字列とを対に有する異表記パターンを１以上格納しており、コンピュータを、１以上の用語を受け付ける受付部と、前記受付部が受け付けた１以上の各用語に対して、前記記憶媒体の１以上の各異表記パターンを適用し、１以上の用語を生成し、前記１以上の各用語と前記生成した用語とを有する１以上の異表記の候補の用語対である異表記候補用語対を生成する用語対生成部と、前記用語対生成部が生成した１以上の異表記候補用語対ごとに、字種関連素性、辞書関連素性、類似度素性のうちの一以上の素性を含む複数の素性を取得する素性取得部と、前記用語対生成部が生成した各異表記候補用語対に対して、前記記憶媒体の２以上の学習データと、前記素性取得部が取得した複数の素性とを用いて、教師あり機械学習法により、前記用語対格納部の各異表記候補用語対が異表記の用語対であるか否かを判断する機械学習部と、前記機械学習部における判断結果を出力する出力部として機能させるためのプログラム、である。

また、図２２は、本明細書で述べたプログラムを実行して、上述した実施の形態の異表記取得装置等を実現するコンピュータの外観を示す。上述の実施の形態は、コンピュータ

ハードウェア及びその上で実行されるコンピュータプログラムで実現され得る。図２２は、このコンピュータシステム３４０の概観図であり、図２３は、コンピュータシステム３４０のブロック図である。

【０３２６】

図２２において、コンピュータシステム３４０は、ＦＤドライブ、ＣＤ－ＲＯＭドライブを含むコンピュータ３４１と、キーボード３４２と、マウス３４３と、モニタ３４４とを含む。

【０３２７】

図２３において、コンピュータ３４１は、ＦＤドライブ３４１１、ＣＤ－ＲＯＭドライブ３４１２に加えて、ＭＰＵ３４１３と、ＣＤ－ＲＯＭドライブ３４１２及びＦＤドライブ３４１１に接続されたバス３４１４と、ブートアッププログラム等のプログラムを記憶するためのＲＯＭ３４１５とに接続され、アプリケーションプログラムの命令を一時的に記憶するとともに一時記憶空間を提供するためのＲＡＭ３４１６と、アプリケーションプログラム、システムプログラム、及びデータを記憶するためのハードディスク３４１７とを含む。ここでは、図示しないが、コンピュータ３４１は、さらに、ＬＡＮへの接続を提供するネットワークカードを含んでも良い。

【０３２８】

コンピュータシステム３４０に、上述した実施の形態の異表記取得装置等の機能を実行させるプログラムは、ＣＤ－ＲＯＭ３５０１、またはＦＤ３５０２に記憶されて、ＣＤ－ＲＯＭドライブ３４１２またはＦＤドライブ３４１１に挿入され、さらにハードディスク３４１７に転送されても良い。これに代えて、プログラムは、図示しないネットワークを介してコンピュータ３４１に送信され、ハードディスク３４１７に記憶されても良い。プログラムは実行の際にＲＡＭ３４１６にロードされる。プログラムは、ＣＤ－ＲＯＭ３５０１、ＦＤ３５０２またはネットワークから直接、ロードされても良い。

【０３２９】

プログラムは、コンピュータ３４１に、上述した実施の形態の異表記取得装置等の機能を実行させるオペレーティングシステム（ＯＳ）、またはサードパーティープログラム等は、必ずしも含まなくても良い。プログラムは、制御された態様で適切な機能（モジュール）を呼び出し、所望の結果が得られるようにする命令の部分のみを含んでいれば良い。コンピュータシステム３４０がどのように動作するかは周知であり、詳細な説明は省略する。

【０３３０】

また、上記プログラムを実行するコンピュータは、単数であってもよく、複数であってもよい。すなわち、集中処理を行ってもよく、あるいは分散処理を行ってもよい。

【０３３１】

また、上記各実施の形態において、一の装置に存在する２以上の通信手段は、物理的に一の媒体で実現されても良いことは言うまでもない。

【０３３２】

また、上記各実施の形態において、各処理（各機能）は、単一の装置（システム）によって集中処理されることによって実現されてもよく、あるいは、複数の装置によって分散処理されることによって実現されてもよい。

【０３３３】

本発明は、以上の実施の形態に限定されることなく、種々の変更が可能であり、それらも本発明の範囲内に包含されるものであることは言うまでもない。

【産業上の利用可能性】

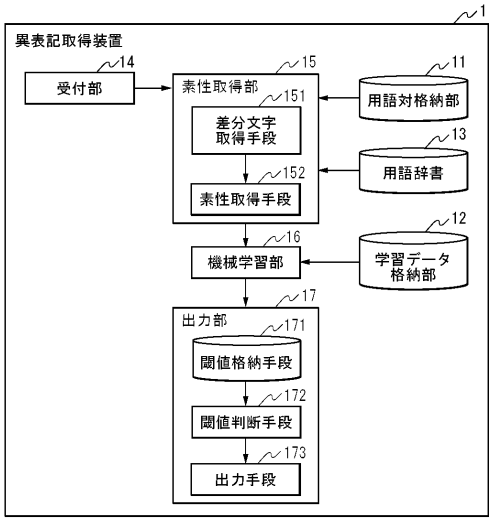
【０３３４】

以上のように、本発明にかかる異表記取得装置は、用語対の分野を問わず、精度の高い異表記の用語対の抽出が可能となる、という効果を有し、異表記取得装置等として有用である。

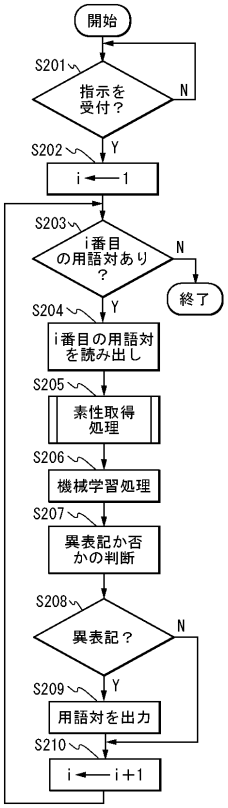
【符号の説明】

【０３３５】
１、２　異表記取得装置
１１　用語対格納部
１２　学習データ格納部
１３　用語辞書
１４、２３　受付部
１５　素性取得部
１６　機械学習部
１７　出力部
２１　異表記用語対格納部
２２　異表記パターン格納部
２４　編集箇所取得部
２５　異表記パターン取得部
２６　異表記パターン蓄積部
２７　用語対生成部
１５１　差分文字取得手段
１５２　素性取得手段
１７１　閾値格納手段
１７２　閾値判断手段
１７３　出力手段

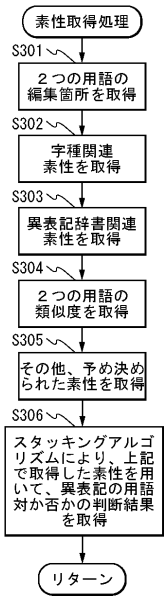
【図１】



【図２】



【図３】

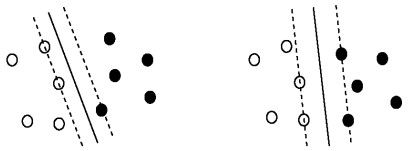


【図４】

用語	読み	品詞		代表表記	カテゴリ	ドメイン	.....
子ども	こども	名詞	普通名詞	子供／こども	人	――	――
リンゴ	りんご	名詞	普通名詞	林檎／りんご	植物；人工物-食べ物	料理・食事	――
おくれた	おくれた	動詞	母音動詞	送れる／おくれる	――	――	――
.....	.....	.....	.....	.....	.....	.....	.....

【図５】

（A）スモールマージン　（B）ラージマージン



【図６】

日本語用語対	クローズドデータにおける割合	オープンデータにおける割合
日本語異表記対である	5.25%　(745/14185)	5.11%　(725/14185)
日本語異表記対でない	94.74%（13440/14185）	94.88%（13460/14185）

【図７】

一致度Ｋの値	一致度の評価
$0 \leq K < 0.2$	ごく軽度の一致
$0.2 \leq K < 0.4$	軽度の一致
$0.4 \leq K < 0.6$	中等度の一致
$0.6 \leq K < 0.8$	高度の一致
$0.8 \leq K$	ほぼ完全な一致

【図８】

手法	10ＣＶ		ＯＣ	
	正解率	Ｆ値	正解率	Ｆ値
ベースライン	96.41%	0.483	96.36%	0.452
全素性	99.15%	0.916	99.12%	0.912

【図９】

省いた素性	10CV		OC		省いた素性	10CV		OC		省いた素性	10CV		OC	
	全素性	素性選択	全素性	素性選択		全素性	素性選択	全素性	素性選択		全素性	素性選択	全素性	素性選択
S1	1789	7809	7775	1456	S27	0	6305	6021	3474	S53	9361	630	6206	3424
S2	4585	4418	8652	0	S28	0	8651	0	0	S54	9462	392	9682	275
S3	3918	5810	9018	847	S29	2805	6079	3892	4083	S55	9996	4	9942	57
S4	3390	5869	7688	2302	S30	5639	4167	5292	3917	S56	6061	1828	0	0
S5	6081	2803	3388	5243	S31	2486	5638	3936	5844	S57	0	9533	0	0
S6	0	9933	4131	5694	S32	2816	5766	6410	0	S58	10000	0	10000	0
S7	6352	0	0	0	S33	4740	4750	2230	7769	S59	0	0	6314	0
S8	5203	4477	1810	6978	S34	8022	1770	3375	6123	S60	0	0	0	0
S9	5150	4294	777	9196	S35	8045	1781	8661	0	S61	5062	4629	215	9509
S10	285	9376	128	9761	S36	4773	4799	5283	3910	S62	6086	1841	6302	0
S11	3504	6156	5848	3957	S37	0	0	0	0	S63	8658	0	0	6353
S12	4272	5631	7275	2414	S38	0	0	3488	6006	S64	3463	3403	0	0
S13	9246	469	8754	772	S39	6021	1832	3422	6079	S65	0	0	0	0
S14	6080	2802	6320	0	S40	7670	1838	8696	0	S66	0	0	6341	0
S15	3480	3408	6355	0	S41	8623	0	6325	0	S67	9999	1	10000	0
S16	3505	3423	6292	0	S42	489	9021	1854	7692	S68	8075	1859	1024	8973
S17	3862	4009	8610	0	S43	0	9501	3167	6402					
S18	3468	3499	6319	0	S44	9993	4	1230	8505					
S19	0	6217	0	0	S45	6290	0	0	0					
S20	1780	6126	0	0	S46	0	0	0	0					
S21	0	0	3477	5158	S47	9845	0	2114	7708					
S22	6137	1818	6380	0	S48	0	0	9496	0					
S23	0	6324	6010	3497	S49	3429	3490	3514	3394					
S24	0	0	0	0	S50	7686	954	9489	0					
S25	6101	1746	8686	0	S51	0	9790	6304	0					
S26	3983	3901	6366	0	S52	0	8678	9546	0					

【図１０】

番号	素性の種類	番号	素性の種類
S1	一つ目の表記の編集箇所	S45	S1の品詞と位置情報
S2	二つ目の表記の編集箇所	S46	S2の品詞と位置情報
S3	編集箇所の前方の1文字	S47	S3の品詞と位置情報
S4	編集箇所の後方の1文字	S48	S4の品詞と位置情報
S5	編集箇所の前方の2文字連続	S49	S13の品詞と位置情報
S6	編集箇所の前方の3文字連続	S50	S14の品詞と位置情報
S7	編集箇所の前方2文字目の文字	S51	S17の品詞と位置情報
S8	編集箇所の前方3文字目の文字	S52	S18の品詞と位置情報
S9	編集箇所の後方の2文字	S53	日本語単語対の類似度
S10	編集箇所の後方の3文字	S54	スタッキングアルゴリズムを使用して、日本語用語対のJUMAN辞書の代表表記が一致するかどうか
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		
S13	'S1の情報_S2の情報' とした文字列	S55	日本語用語対の文字数が同数で編集箇所が両方とも数字の場合、同じ値か違う値かどうか
S14	'S3の情報_S13の情報' とした文字列		
S15	'S5の情報_S13の情報' とした文字列		
S16	'S6の情報_S13の情報' とした文字列	S56	日本語用語対の文字数が同数で編集箇所が両方ともひらがなの場合、同じ音声か違う音声かどうか
S17	'S13の情報_S4の情報' とした文字列		
S18	'S3の情報_S13の情報_S4の情報' とした文字列	S57	日本語用語対の文字数が同数で編集箇所が両方ともカタカナの場合、同じ音声か違う音声かどうか
S19	'S5の情報_S13の情報_S4の情報' とした文字列		
S20	'S6の情報_S13の情報_S4の情報' とした文字列	S58	日本語用語対の文字数が同数で編集箇所が両方ともローマ字の場合、大文字と小文字の違いだけかどうか
S21	'S13の情報_S7の情報' とした文字列	S59	日本語用語対の文字数が同数で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かどうか
S22	'S3の情報_S13の情報_S7の情報' とした文字列		
S23	'S5の情報_S13の情報_S7の情報' とした文字列	S60	日本語用語対の文字数が同数で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かどうか
S24	'S6の情報_S13の情報_S7の情報' とした文字列		
S25	'S13の情報_S8の情報' とした文字列	S61	編集箇所が日本語用語対の一方にしかなく、その編集箇所が「化」、「系」、「類」、「型」、「形」、「氏」、「ー」、「・」かどうか
S26	'S3の情報_S13の情報_S8の情報' とした文字列		
S27	'S5の情報_S13の情報_S8の情報' とした文字列	S62	編集箇所が日本語用語対の一方にしかなく、その編集箇所の単語が日本語用語対の最後の文字と一致するかどうか
S28	'S6の情報_S13の情報_S8の情報' とした文字列		
S29	S1の字種	S63	編集箇所が日本語用語対の一方にしかなく、その編集箇所が桁数をあらわす用語かどうか(例.「千」,「万」)
S30	S2の字種		
S31	S3の字種		
S32	S4の字種	S64	日本語用語対のJUMAN辞書の定義されている代表表記が一致するかどうか
S33	S13の字種	S65	日本語用語対が日本語ワードネット辞書に類義語として定義されているかどうか
S34	S14の字種		
S35	S17の字種		
S36	S18の字種	S66	日本語用語対の編集箇所が異体字辞書に異体字として定義されているかどうか
S37	S1の品詞		
S38	S2の品詞		
S39	S3の品詞	S67	日本語用語対の文字数が同数で編集箇所が漢字とひらがなの場合、JUMAN辞書の読みが一致するかどうか
S40	S4の品詞		
S41	S13の品詞		
S42	S14の品詞	S68	日本語用語対の文字数が同数で編集箇所が両方とも漢字の場合、JUMAN辞書の読みが一致するかどうか
S43	S17の品詞		
S44	S18の品詞		

【図１１】

辞書	割合
EDR辞書	20.45% (4948/24185)
日本語ワードネット辞書	1.71% (1414/82270)
JUMAN辞書	6.52% (1523/23348)

【図１２】

	辞書		
	EDR辞書	日本語ワードネット辞書	JUMAN辞書
正解率	45.15%(10800/23920)	21.36%(16814/78717)	83.34%(19249/23097)

【図１３】

テストデータ	日本語異表記対に分類されたデータ	非日本語異表記対に分類されたデータ
用語対DB	書いてた頃:書いていた頃 JAVAScript:JAVA Script シュウ:シュウジ 自サーバー:自サーバ 日光彫:日光彫り	仕上塗料:上塗料 約１２ｃｍ:１２ｃｍ ｂｏｙ:ｂｏｙｓ Ｌｖ６０:Ｌｖ６０台 １６０ＧＢ:６０ＧＢ
EDR辞書	竜山文化:龍山文化 そっくり返る:反っくり返る 言い替えする:言替えする 爽かだ:爽快だ ローテーション:ローテイション	前:前方 通詞:通辞 極道だ:獄道だ 科目:目 仮想専用線網:仮想専用網
日本語ワードネット辞書	寄せ合せ:寄せ合わせ 追尋:追従 跳ねる:跳返る 起きあがる:起き上がる ご存じ:ご存知	異母兄弟:異父兄弟 若い者:若い衆 凶漢:悪漢 貴:高貴 切り:切間
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 すり粉木:擂り粉木 けいし:けい視 りょうち:りょう地	洞々たる:洞洞たる やけ残る:焼け残る 糯米:餅米 しがい:市がい 書き出す:書出す 奇れいすきだ:綺れいすきだ

【図１６】

	再現率	適合率	総数	0	1
0	99.97	96.35	26900	26892	1018
1	30.75	98.26	1470	8	452
総数	96.38	96.38	28370	26900	1470
F値 0: 0.9813 1: 0.4684					

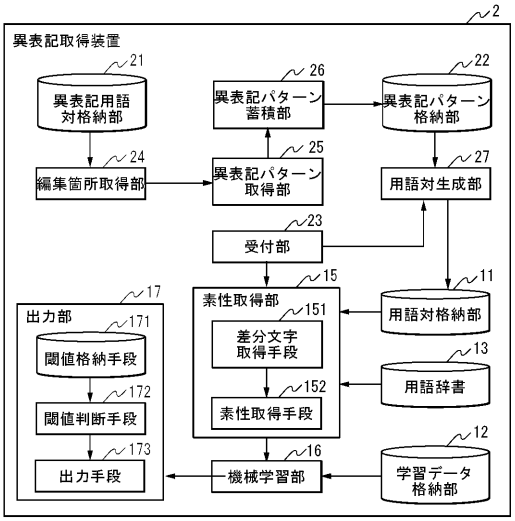
【図１７】

	再現率	適合率	総数	0	1
0	99.99	96.36	13440	13438	507
1	31.95	99.17	745	2	238
総数	96.41	96.41	14185	13440	745
F値 0: 0.9814 1: 0.4833					

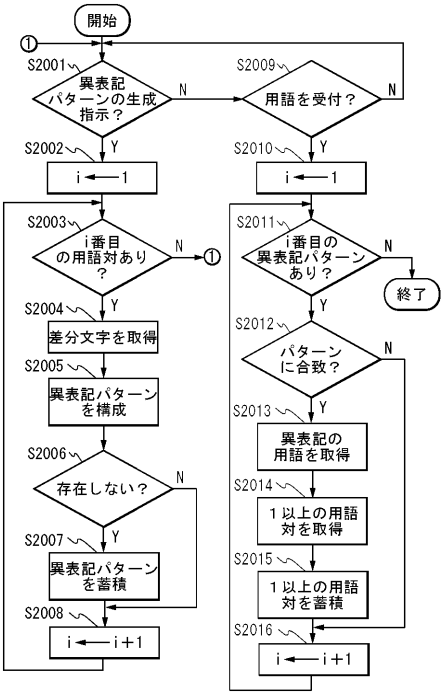
【図１８】

	再現率	適合率	総数	0	1
0	99.96	96.34	13460	13454	511
1	29.52	97.27	725	6	214
総数	96.36	96.36	14185	13460	725
F値 0: 0.9812 1: 0.4529					

【図１９】



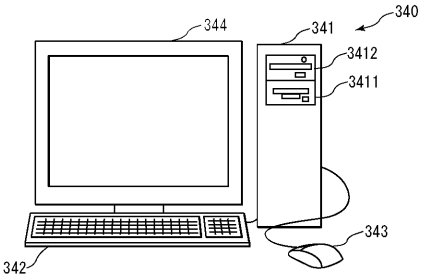
【図２０】



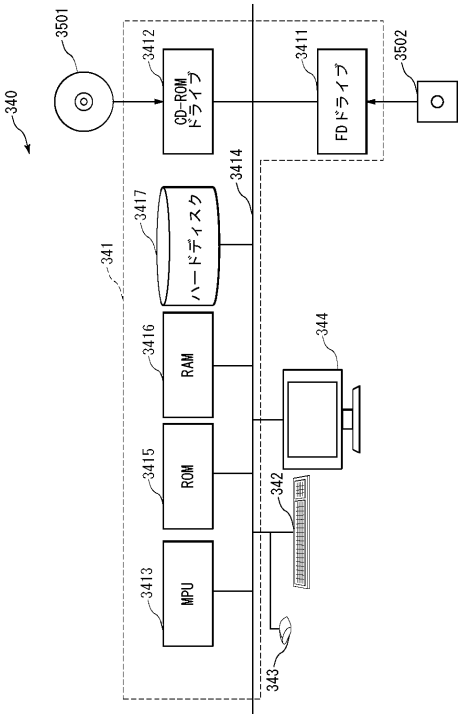
【図２１】

異表記パターン	
del	・
del	―
del	い
―	1
イ	ィ
取	と
当	あ
言	い
⋮	⋮

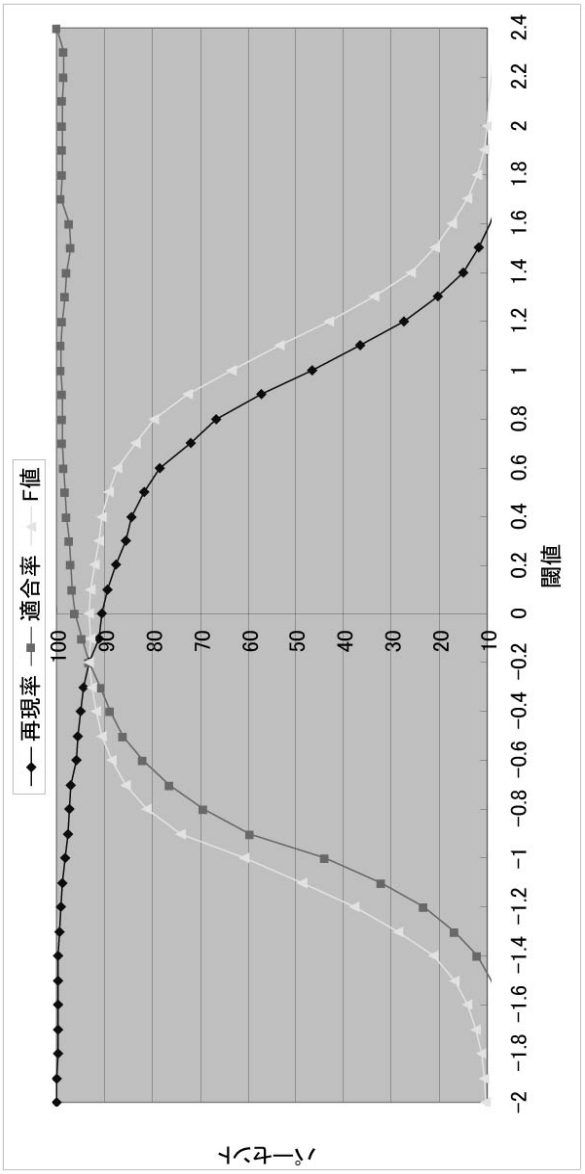
【図２２】



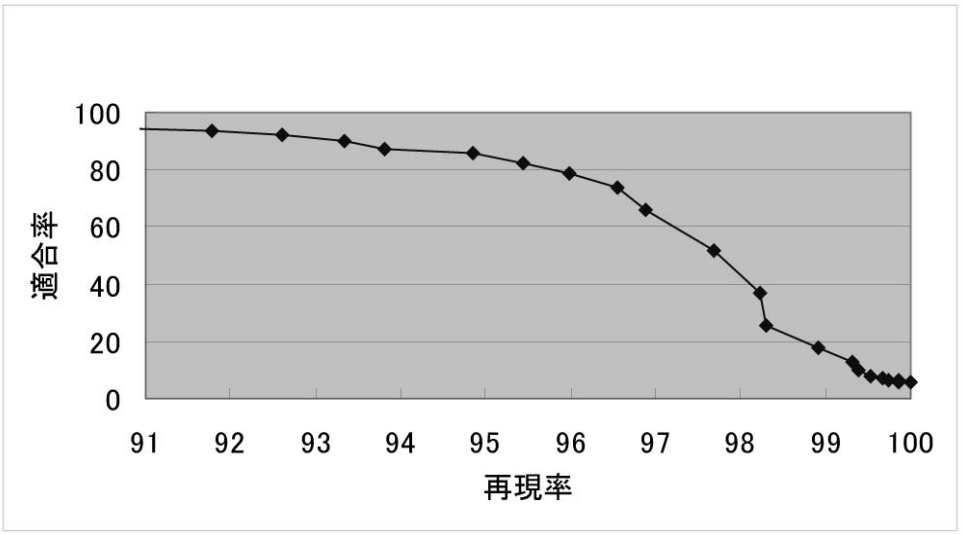
【図２３】



【図１４】


再現率
適合率
F値
閾値
100
90
80
70
60
50
40
30
20
10
2.4
2.2
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
−1.2
−1.4
−1.6
−1.8
−2

【図１５】


適合率
再現率
100
80
60
40
20
0
91
92
93
94
95
96
97
98
99
100

フロントページの続き

(72)発明者　風間　淳一
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　黒田　航
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内
(72)発明者　藤田　篤
東京都小金井市貫井北町４－２－１　独立行政法人情報通信研究機構内

Ｆターム(参考) 5B075 ND03　QM08　UU01
5B091 AA15　AB17　CC01　CC16