

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第5470620号
(P5470620)

(45) 発行日 平成26年4月16日(2014.4.16)

(24) 登録日 平成26年2月14日(2014.2.14)

(51) Int.Cl. F I
G O 6 F 17/27 (2006.01) G O 6 F 17/27 Z

請求項の数 10 (全 48 頁)

(21) 出願番号	特願2009-299287 (P2009-299287)	(73) 特許権者	301022471
(22) 出願日	平成21年12月30日(2009.12.30)		独立行政法人情報通信研究機構
(65) 公開番号	特開2011-138440 (P2011-138440A)		東京都小金井市貫井北町4-2-1
(43) 公開日	平成23年7月14日(2011.7.14)	(74) 代理人	100115749
審査請求日	平成24年11月19日(2012.11.19)		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) 【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子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であって、
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、

10

20

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同じ値の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。

10

【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。

20

【請求項４】

前記用語情報は、

前記用語に対応付けて、当該用語の読みをも有し、

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。

30

【請求項５】

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４いずれか記載の異表記取得装置。

【請求項６】

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、

40

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５記載の異表記取得装置。

【請求項７】

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１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ことを示す情報である正負情報とを対

50

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
各異表記パターンが有するいずれかの文字列を含むか否かを判断し、含むと判断した場合
に、当該用語に含まれる当該文字列の部分を、当該異表記パターンの文字列であり、当該
含まれる文字列の他方の文字列に置き換えて、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であって
、

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、異表記
の用語対ごとに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、異表記の用語対ごとに、当該用語対を構成
する一の用語の文字列に含まれる編集箇所を第一文字列とし、当該用語対を構成する他の
用語の文字列に含まれる編集箇所を第二文字列として取得し、当該第一文字列と当該第二
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異表記取得装置。

20

【請求項8】

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7いずれか記載の異表記取得装置。

30

40

【請求項9】

記憶媒体に、
編集距離が1以上の用語対、
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、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
記の用語対であること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、

50

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であって、

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、

前記素性取得ステップにおいて、

前記記憶媒体の用語対ごとに、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異表記取得方法。

【請求項10】

記憶媒体に、

編集距離が1以上の用語対、

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、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って、

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ものとして、コンピュータを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】

10

20

30

40

50

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。

10

【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.

20

【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-siryuu-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.

30

【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子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1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こと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1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。

40

【0010】

50

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。

【 0 0 1 1 】

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。

【 0 0 1 2 】

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。

【 0 0 1 3 】

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。

【 0 0 1 4 】

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。

【 0 0 1 5 】

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。

【 0 0 1 6 】

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。

【 0 0 1 7 】

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。

【 0 0 1 8 】

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。

【 0 0 1 9 】

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。

10

20

30

40

50

【 0 0 2 0 】

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。

【 0 0 2 1 】

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。

【 0 0 2 2 】

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。

10

【 0 0 2 3 】

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。

【 0 0 2 4 】

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。

20

【 0 0 2 5 】

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こと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。

30

【 0 0 2 6 】

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。

40

【 0 0 2 7 】

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。

【 0 0 2 8 】

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

50

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。

10

【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】

20

【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

30

【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

40

【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】

50

【 0 0 3 3 】

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。

【 0 0 3 4 】

(実施の形態 1)

【 0 0 3 5 】

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。

10

【 0 0 3 6 】

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 1 1、学習データ格納部 1 2、用語辞書 1 3、受付部 1 4、素性取得部 1 5、機械学習部 1 6、出力部 1 7 を備える。素性取得部 1 5 は、差分文字取得手段 1 5 1、素性取得手段 1 5 2 を備える。出力部 1 7 は、閾値格納手段 1 7 1、閾値判断手段 1 7 2、出力手段 1 7 3 を備える。

【 0 0 3 7 】

用語対格納部 1 1 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。

20

【 0 0 3 8 】

学習データ格納部 1 2 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 1 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。

【 0 0 3 9 】

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。

30

【 0 0 4 0 】

また、辞書関連素性とは、用語辞書 1 3 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 1 6 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

40

50

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。2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1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「三者会談」「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】

10

20

30

40

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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。

10

【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」「肝炎ウイルス」(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」「ー」を取得する。

【0061】

20

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1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子種が異なり、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。

30

40

【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

50

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。

10

【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。

20

【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つのスコアのうちの、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つた。

30

40

【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

50

コアの小さい方を取得し、また、「I」「J」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」「J」のスコアが0に近い方を取得し、また、「I」「J」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。

10

【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。

20

【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。

30

【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るても含まないと考えるても良い。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。

40

【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。

50

【 0 0 8 3 】

(ステップS 2 0 2) 素性取得部 1 5 は、カウンタ i に 1 を代入する。

【 0 0 8 4 】

(ステップS 2 0 3) 素性取得部 1 5 は、 i 番目の用語対が用語対格納部 1 1 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S 2 0 4 に行き、存在しなければ処理を終了する。

【 0 0 8 5 】

(ステップS 2 0 4) 素性取得部 1 5 は、用語対格納部 1 1 から、 i 番目の用語対を読み出す。

【 0 0 8 6 】

(ステップS 2 0 5) 素性取得部 1 5 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。

【 0 0 8 7 】

(ステップS 2 0 6) 機械学習部 1 6 は、ステップS 2 0 5 で取得された複数の素性と、学習データ格納部 1 2 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。

【 0 0 8 8 】

(ステップS 2 0 7) 出力部 1 7 は、ステップS 2 0 6 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 1 7 を構成する閾値判断手段 1 7 2 は、閾値格納手段 1 7 1 から閾値を読み出し、ステップS 2 0 6 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。

【 0 0 8 9 】

(ステップS 2 0 8) 出力部 1 7 は、ステップS 2 0 8 での判断結果が、異表記の用語対であればステップS 2 0 9 に行き、異表記の用語対でなければステップS 2 1 0 に行く。

【 0 0 9 0 】

(ステップS 2 0 9) 出力部 1 7 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。

【 0 0 9 1 】

(ステップS 2 1 0) 素性取得部 1 5 は、カウンタ i を 1 , インクリメントする。ステップS 2 0 3 に戻る。

【 0 0 9 2 】

次に、ステップS 2 0 5 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。

【 0 0 9 3 】

(ステップS 3 0 1) 素性取得部 1 5 を構成する差分文字取得手段 1 5 1 は、2 つの用語の編集箇所を取得する。

【 0 0 9 4 】

(ステップS 3 0 2) 素性取得部 1 5 の素性取得手段 1 5 2 は、ステップS 3 0 1 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。

【 0 0 9 5 】

(ステップS 3 0 3) 素性取得手段 1 5 2 は、用語辞書 1 3 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。

【 0 0 9 6 】

(ステップS 3 0 4) 素性取得手段 1 5 2 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。

【 0 0 9 7 】

10

20

30

40

50

(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、「ヨ」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。

10

20

30

40

50

【 0 1 0 8 】

S 7 は、「編集箇所前方 2 文字目の文字」であり、上記具体例では、「シ」である。

【 0 1 0 9 】

S 8 は「編集箇所前方 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。

【 0 1 1 0 】

S 9 は「編集箇所後方の 2 文字」であり、上記具体例では、「ウイ」である。

【 0 1 1 1 】

S 1 0 は「編集箇所後方の 3 文字」であり、上記具体例では、「ウイン」である。

【 0 1 1 2 】

S 1 1 は「編集箇所後方 2 文字目の文字」であり、上記具体例では、「イ」である。

【 0 1 1 3 】

S 1 2 は「編集箇所後方 3 文字目の文字」であり、上記具体例では、「ン」である。

【 0 1 1 4 】

S 1 3 は「' S 1 の情報 - S 2 の情報' としての文字列」であり、上記具体例では、「ウ -
ー」である。

【 0 1 1 5 】

S 1 4 は「' S 3 の情報 - S 1 3 の情報' としての文字列」であり、上記具体例では、「ヨ
- ウ - ー」である。

【 0 1 1 6 】

S 1 5 は「' S 5 の情報 - S 1 3 の情報' としての文字列」であり、上記具体例では、「シ
ヨ - ウ - ー」である。

【 0 1 1 7 】

S 1 6 は「' S 6 の情報 - S 1 3 の情報' としての文字列」であり、上記具体例では、「シ
ヨ - ウ - ー」である。

【 0 1 1 8 】

S 1 7 は「' S 1 3 の情報 - S 4 の情報' 」であり、上記具体例では、「ウ - ー - ウ」で
ある。

【 0 1 1 9 】

S 1 8 は「' S 3 の情報 - S 1 3 の情報 - S 4 の情報' 」であり、上記具体例では、「ヨ
- ウ - ー - ウ」である。

【 0 1 2 0 】

S 1 9 は「' S 5 の情報 - S 1 3 の情報 - S 4 の情報' としての文字列」であり、上記具体
例では、「シヨ - ウ - ー - ウ」である。

【 0 1 2 1 】

S 2 0 は「' S 6 の情報 - S 1 3 の情報 - S 4 の情報' としての文字列」であり、上記具体
例では、「シヨ - ウ - ー - ウ」である。

【 0 1 2 2 】

S 2 1 は「' S 1 3 の情報 - S 7 の情報' としての文字列」であり、上記具体例では、「ウ
- ー - シ」である。

【 0 1 2 3 】

S 2 2 は「' S 3 の情報 - S 1 3 の情報 - S 7 の情報' としての文字列」であり、上記具体
例では、「ヨ - ウ - ー - シ」である。

【 0 1 2 4 】

S 2 3 は「' S 5 の情報 - S 1 3 の情報 - S 7 の情報' としての文字列」であり、上記具体
例では、「シヨ - ウ - ー - シ」である。

【 0 1 2 5 】

S 2 4 は「' S 6 の情報 - S 1 3 の情報 - S 7 の情報' としての文字列」であり、上記具体
例では、「シヨ - ウ - ー - シ」である。

【 0 1 2 6 】

10

20

30

40

50

S 2 5 は「' S 1 3 の情報 - S 8 の情報' としての文字列」であり、上記具体例では、「ウ
- - - (d e l)」である。

【 0 1 2 7 】

S 2 6 は「' S 3 の情報 - S 1 3 の情報 - S 8 の情報' としての文字列」であり、上記具体例では、「ヨ - ウ - - - (d e l)」である。

【 0 1 2 8 】

S 2 7 は「' S 5 の情報 - S 1 3 の情報 - S 8 の情報' としての文字列」であり、上記具体例では、「シヨ - ウ - - - (d e l)」である。

【 0 1 2 9 】

S 2 8 は「' S 6 の情報 - S 1 3 の情報 - S 8 の情報' としての文字列」であり、上記具体例では、「シヨ - ウ - - - (d e l)」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 1 5 2 は、S 3 から S 2 8 の素性を取得できる。

10

【 0 1 3 0 】

S 2 9 は「S 1 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。

【 0 1 3 1 】

S 3 0 は「S 2 の字種」であり、上記具体例では、「カタカナ」である。

【 0 1 3 2 】

S 3 1 は「S 3 の字種」であり、上記具体例では、「カタカナ」である。

20

【 0 1 3 3 】

S 3 2 は「S 4 の字種」であり、上記具体例では、「カタカナ」である。

【 0 1 3 4 】

S 3 3 は「S 1 3 の字種」であり、上記具体例では、「カタカナ」である。

【 0 1 3 5 】

S 3 4 は「S 1 4 の字種」であり、上記具体例では、「カタカナ」である。

【 0 1 3 6 】

S 3 5 は「S 1 7 の字種」であり、上記具体例では、「カタカナ」である。

【 0 1 3 7 】

S 3 6 は「S 1 8 の字種」であり、上記具体例では、「カタカナ」である。

30

【 0 1 3 8 】

S 3 7 は「S 1 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 S 1 ）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。

【 0 1 3 9 】

S 3 8 は「S 2 の品詞」であり、上記具体例では、「名詞」である。

【 0 1 4 0 】

S 3 9 は「S 3 の品詞」であり、上記具体例では、「名詞」である。

40

【 0 1 4 1 】

S 4 0 は「S 4 の品詞」であり、上記具体例では、「名詞」である。

【 0 1 4 2 】

S 4 1 は「S 1 3 の品詞」であり、上記具体例では、「名詞」である。

【 0 1 4 3 】

S 4 2 は「S 1 4 の品詞」であり、上記具体例では、「名詞」である。

【 0 1 4 4 】

S 4 3 は「S 1 7 の品詞」であり、上記具体例では、「名詞」である。

【 0 1 4 5 】

S 4 4 は「S 1 8 の品詞」であり、上記具体例では、「名詞」である。

50

【 0 1 4 6 】

S 4 5 は「 S 1 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。

【 0 1 4 7 】

S 4 6 は「 S 2 の品詞と位置情報」であり、上記具体例では、「名詞， 3」である。

【 0 1 4 8 】

S 4 7 は「 S 3 の品詞と位置情報」であり、上記具体例では、「名詞， 2」である。

【 0 1 4 9 】

S 4 8 は「 S 4 の品詞と位置情報」であり、上記具体例では、「名詞， 4」である。

【 0 1 5 0 】

S 4 9 は「 S 1 3 の品詞と位置情報」であり、上記具体例では、「名詞， 3」である。

10

【 0 1 5 1 】

S 5 0 は「 S 1 4 の品詞と位置情報」であり、上記具体例では、「名詞， 2」である。

【 0 1 5 2 】

S 5 1 は「 S 1 7 の品詞と位置情報」であり、上記具体例では、「名詞， 6」である。

【 0 1 5 3 】

S 5 2 は「 S 1 8 の品詞と位置情報」であり、上記具体例では、「名詞， 3」である。

【 0 1 5 4 】

S 5 3 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。

【 0 1 5 5 】

S 5 4 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 J U M A N 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 1 6 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、 J U M A N 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、 J U M A N 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、 9 0 4 6 1 2 組、取り出せる。そのうち、代表表記が等しい単語対 (2 5 9 3 4 組) を取り出す。次に、 J U M A N 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 (S 1 から S 6 8 の素性) のうち、 S 5 4 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、 S 5 4 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 S 5 4 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、 J U M A N 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 J U M A N 辞書に記載されていない用語対に対しても、 J U M A N 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。

20

30

【 0 1 5 6 】

S 5 5 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、 S 5 5 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、 S 5 5 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。

40

【 0 1 5 7 】

S 5 6 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、「おかあちゃん」「おかあちゃん」の用語対の場合は、 S 5 6 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、 S 5 6 は、「日本語用語対の編集箇所が両方ともひらがな

50

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」「オリーブ・オイル」の用語対の場合、「ウインドウ」「ウインドウ」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、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。

10

【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」「触れるくらい」の用語対の場合、「飲むくらい」「飲むくらい」の用語対の場合、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。

20

【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、S61の素性は「1」となる。

30

【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、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かず用語かどうか（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、S63の素性は「1」となる。

40

【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】

50

S 6 5 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 1 5 2 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。

【 0 1 6 7 】

S 6 6 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。

【 0 1 6 8 】

S 6 7 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。

【 0 1 6 9 】

S 6 8 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。

【 0 1 7 0 】

また、上記の 6 8 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。

【 0 1 7 1 】

また、上記の 6 8 の素性をグループ化すると、例えば、以下の G 1 から G 7 のグループに分かれる、と考えられる。

【 0 1 7 2 】

G 1 は、S 1 から S 5 2 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。

【 0 1 7 3 】

G 2 は、S 5 3 の素性であり、類似度素性である。

【 0 1 7 4 】

G 3 は、S 5 4 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。

【 0 1 7 5 】

G 4 は、S 5 5 から S 6 0 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。

【 0 1 7 6 】

G 5 は、S 6 1 から S 6 3 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。

【 0 1 7 7 】

G 6 は、S 6 4 から S 6 6 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。

【 0 1 7 8 】

G 7 は、S 6 7 から S 6 8 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。

【 0 1 7 9 】

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 1 2 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 1 2 に格納しておく。

10

20

30

40

50

【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。

10

20

30

40

50

【 0 1 9 4 】

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。

【 0 1 9 5 】

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 】

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

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

10

【 0 1 9 6 】

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 ($F, 1 \leq j \leq k$) 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、スムージングを行う。ここでは、以下の数式 2 を用いてスムージングを行ったものを用いる。

20

【 数 2 】

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

【 0 1 9 7 】

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

30

【 0 1 9 8 】

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を与えられたときに、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。

【 0 1 9 9 】

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 ($F, 1 \leq j \leq k$) 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 】

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

40

【 0 2 0 0 】

ただし、 f_{\max} は以下の数式 4 によって与えられる。

【 数 4 】

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

【 0 2 0 1 】

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。

50

【 0 2 0 2 】

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 ($1 \leq j \leq k$) 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】

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

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

10

【数 6】

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

【 0 2 0 3 】

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。

【 0 2 0 4 】

20

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。

【 0 2 0 5 】

参考文献 (1) : Eric Sven Ristad, Maximum Entropy Modeling for Natural Language, (ACL/EACL Tutorial Program, Madrid, 1997)

【 0 2 0 6 】

30

参考文献 (2) : Eric Sven Ristad, Maximum Entropy Modeling Toolkit, Release 1.6 beta, (<http://www.mnemonic.com/software/memtool>, 1998)

【 0 2 0 7 】

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。

【 0 2 0 8 】

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。

40

【 0 2 0 9 】

このとき、二つの分類が正例と負例からなるものとする、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。

【 0 2 1 0 】

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。

50

【 0 2 1 1 】

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 】

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

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

10

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

【 0 2 1 2 】

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 ($i = 1, \dots, l, y_j \in \{1, -1\}$) は学習データの文脈と分類先を意味し、関数 sgn は、

$$\text{sgn}(x) = \begin{cases} 1 & (x \geq 0) \\ -1 & (\text{otherwise}) \end{cases}$$

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。

20

【 数 8 】

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

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

$$\sum_{j=1}^l \alpha_j y_j = 0 \quad \text{式 (8-3)}$$

【 0 2 1 3 】

30

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 】

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

【 0 2 1 4 】

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。

40

【 0 2 1 5 】

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。

【 0 2 1 6 】

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)

【 0 2 1 7 】

参考文献 (4) : Taku Kudoh, Tinsvsm: Support Vector machines, (<http://cl.aist.nar>

50

a.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。

10

【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。

20

【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。

30

【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。

40

【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

50

、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。

10

【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。

20

（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。

30

【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 T S U B A K I (<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、カップパ統計量 K を用いて判定する。

40

【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

50

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。

10

【0233】

また、クロ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、オープンクロスによる学習により精度の測定を行う。

20

【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。

30

【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。

40

(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ができ、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】

50

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。

10

【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ができる。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。

20

【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。

30

【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から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。

40

50

【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。

10

【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。

20

【数10】

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

【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。

30

【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個すべてを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。

40

【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

50

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。

10

【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。

20

【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な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。

30

40

（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。

50

【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。

10

【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。

20

（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

30

? 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。

40

（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（

50

S 5 5、S 5 8、S 6 7) が y e s と判定されたものを正例、すべて n o と判定されたものを負例として F 値を求める。

【 0 2 6 7 】

図 1 6 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 1 0 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 1 6 において、「 0 」は負例、「 1 」は正例である。また、図 1 6 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 1 6 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 2 6 8 9 2、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 1 0 1 8、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 4 5 2 であった。また、負例(「 0 」)の再現率は 9 9 . 9 7 %、適合率は 9 6 . 3 5 % であった。また、正例(「 1 」)の再現率は 3 0 . 7 5 %、適合率は 9 8 . 2 6 % であった。また、すべてのデータの再現率は 9 6 . 3 8 %、適合率は 9 6 . 3 8 % 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 1 0 に代入して、算出した負例の F 値は 0 . 9 8 1 3、正例の F 値は 0 . 4 6 8 4 であった。なお、図 1 7 から図 2 1 の各データの意味は、図 1 6 と同様であるので説明を省略する。

10

【 0 2 6 8 】

図 1 7 は、ベースライン手法で、クローズドデータのみを用いて 1 0 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 1 7 において、負例の F 値は 0 . 9 8 1 4、正例の F 値は 0 . 4 8 3 3 であった。

20

【 0 2 6 9 】

図 1 8 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 1 8 において、負例の F 値は 0 . 9 8 1 2、正例の F 値は 0 . 4 5 2 9 であった。

【 0 2 7 0 】

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 . 9 1 2)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)

【 0 2 7 1 】

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「 1 0 0 %」、適合率は「 0 . 0 5 2 5 %」であった。そして、かかる再現率および適合率を、数式 1 0 に代入し、算出された、正例(「 1 」)の F 値は「 0 . 0 9 9 8」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「 9 9 . 1 2 %」、再現率「 9 9 . 0 7 %」、適合率「 9 2 . 2 9 %」、F 値「 0 . 9 1 2」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。

30

40

(応用例)

【 0 2 7 2 】

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 1 4 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 O R) 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 W e b の検索エンジンを起動するだけの処理でも良い。

【 0 2 7 3 】

50

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」となど、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、検索漏れの少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、別のもので考え、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子種が異なり、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

10

20

30

40

50

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つのスコアのうちの、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の、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の、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

10

20

30

40

50

集箇所が2文字以上である場合、1文字ずつの対応とせず、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「編集箇所前方の1文字」「」（なし）、S4「編集箇所後方の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。

10

【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1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1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。

20

30

(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。

40

【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

50

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。

10

【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。

20

【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。

30

【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。

40

【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

50

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。

10

【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。

20

【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「ー」「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「ー」「1」を適用し、「1番目」を生成する。

30

【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「del」「・」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「ー」「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。

40

【0313】

50

(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。

10

【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」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。

20

【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」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。

30

【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、「アイトラッキング」「アイトラッキンぐ」「アイトラッキんぐ」「アイトラッキんいぐ」「アイトラッキんいぐ」「アイトラッキんいぐ」「アイトラッキんいぐ」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」「アイトラッキンぐ」、「アイトラッキング」「アイトラッキんぐ」、「アイトラッキング」「アイトラッキんいぐ」、「アイトラッキング」「アイトラッキんいぐ」、「アイトラッキング」「アイトラッキんいぐ」の7つの用語対を生成する。そして、異表記パターン蓄

40

50

積部 2 6 は、7 つの用語対を異表記パターン格納部 2 2 に蓄積する。

【 0 3 2 0 】

次に、用語対生成部 2 7 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「-」「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。

【 0 3 2 1 】

次に、用語対生成部 2 7 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 2 7 は、「アイトラッキング」を生成する。次に、用語対生成部 2 7 は、用語対「アイトラッキング」「アイトラッキング」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 2 6 は、1 つの用語対を異表記パターン格納部 2 2 に蓄積する。

10

【 0 3 2 2 】

次に、同様に、用語対生成部 2 7 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。

【 0 3 2 3 】

そして、用語対生成部 2 7 は、新たな用語対を用語対格納部 1 1 に蓄積する。

【 0 3 2 4 】

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。

20

【 0 3 2 5 】

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 1 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 1 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。

30

40

また、図 2 2 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 2 2 は、このコンピュータシステム 3 4 0 の概観図であり、図 2 3 は、コンピュータシステム 3 4 0 のブロック図である。

【 0 3 2 6 】

図 2 2 において、コンピュータシステム 3 4 0 は、F D ドライブ、C D - R O M ドライブを含むコンピュータ 3 4 1 と、キーボード 3 4 2 と、マウス 3 4 3 と、モニタ 3 4 4 とを含む。

【 0 3 2 7 】

50

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、MPU 3413 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 ROM 3415 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 RAM 3416 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-ROM 3501、または FD 3502 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 RAM 3416 にロードされる。プログラムは、CD-ROM 3501、FD 3502 またはネットワークから直接、ロードされても良い。

10

【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。

20

【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】

30

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。

40

【符号の説明】

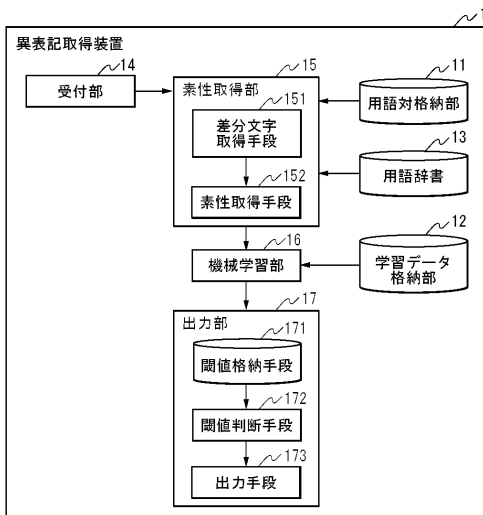
【0335】

- 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

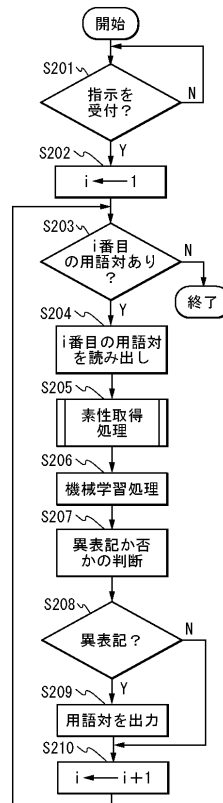
50

- 1 7 出力部
- 2 1 異表記用語対格納部
- 2 2 異表記パターン格納部
- 2 4 編集箇所取得部
- 2 5 異表記パターン取得部
- 2 6 異表記パターン蓄積部
- 2 7 用語対生成部
- 1 5 1 差分文字取得手段
- 1 5 2 素性取得手段
- 1 7 1 閾値格納手段
- 1 7 2 閾値判断手段
- 1 7 3 出力手段

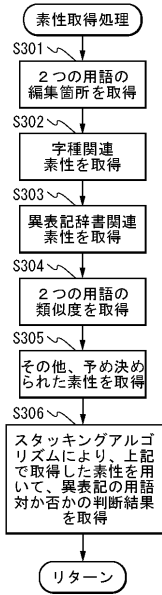
【図1】



【図2】



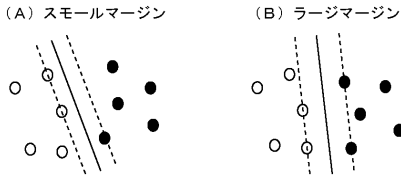
【図3】



【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	子供/こども	人	ドメイン
リンゴ	りんご	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	送れる/おくれる		
.....

【図5】



【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	
	全素性	素性選択	全素性	素性選択
S1	1789	7809	7775	1456
S2	4585	4418	8652	0
S3	3918	5810	9018	847
S4	3390	5869	7688	2302
S5	6081	2803	3388	5243
S6	0	9933	4131	5694
S7	6352	0	0	0
S8	5203	4477	1810	6978
S9	5150	4294	717	9196
S10	285	9376	128	9761
S11	3504	6156	5848	3857
S12	4272	5631	7275	2414
S13	9246	469	8754	772
S14	6080	2802	6320	0
S15	3480	3408	6355	0
S16	3505	3423	6292	0
S17	3862	4009	8610	0
S18	3468	3499	6319	0
S19	0	6217	0	0
S20	1780	6126	0	0
S21	0	0	3477	5158
S22	6137	1818	6380	0
S23	0	6324	6010	3497
S24	0	0	0	0
S25	6101	1746	8686	0
S26	3983	3901	6366	0
省いた素性				
S27				
S28				
S29				
S30				
S31				
S32				
S33				
S34				
S35				
S36				
S37				
S38				
S39				
S40				
S41				
S42				
S43				
S44				
S45				
S46				
S47				
S48				
S49				
S50				
S51				
S52				
省いた素性				
S53				
S54				
S55				
S56				
S57				
S58				
S59				
S60				
S61				
S62				
S63				
S64				
S65				
S66				
S67				
S68				
省いた素性				
S69				
S70				
S71				
S72				
S73				
S74				
S75				
S76				
S77				
S78				
S79				
S80				
S81				
S82				
S83				
S84				
S85				
S86				
S87				
S88				
S89				
S90				
S91				
S92				
S93				
S94				
S95				
S96				
S97				
S98				
S99				
S100				

【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、日本語単語対の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	S55	日本語単語対の文字数が同数で編集箇所が両方も数字の場合、同じ値か違う値かどうか
S12	編集箇所の後方3文字目の文字	S56	日本語単語対の文字数が同数で編集箇所が両方もひらがなの場合、同じ音声が違う音声がどうか
S13	S1の情報_S2の情報とした文字列	S57	日本語単語対の文字数が同数で編集箇所が両方もカタカナの場合、同じ音声が違う音声がどうか
S14	S3の情報_S13の情報とした文字列	S58	日本語単語対の文字数が同数で編集箇所が両方もローマ字の場合、大文字と小文字の違いはどうか
S15	S5の情報_S13の情報とした文字列	S59	日本語単語対の文字数が同数で一方の編集箇所に着点をつけるともう一方の編集箇所になるかどうか
S16	S6の情報_S13の情報とした文字列	S60	日本語単語対の文字数が同数で一方の編集箇所に着点をつけるともう一方の編集箇所になるかどうか
S17	S13の情報_S4の情報とした文字列	S61	編集箇所が日本語単語対の一方にしかなく、その編集箇所が「化、系、類、型、形、氏、一、.、.、.、.」かどうか
S18	S3の情報_S13の情報_S4の情報とした文字列	S62	編集箇所が日本語単語対の一方にしかなく、その編集箇所の単語が日本語単語対の最後の文字と一致するかどうか
S19	S5の情報_S13の情報_S4の情報とした文字列	S63	編集箇所が日本語単語対の一方にしかなく、その編集箇所が桁数をあらず用語かどうか(例:「千、万」)
S20	S6の情報_S13の情報_S4の情報とした文字列	S64	日本語単語対のJUMAN辞書の定義されている代表表記が一致するかどうか
S21	S13の情報_S7の情報とした文字列	S65	日本語単語対が日本語ワードネット辞書に類似語として定義されているかどうか
S22	S3の情報_S13の情報_S7の情報とした文字列	S66	日本語単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
S23	S5の情報_S13の情報_S7の情報とした文字列	S67	日本語単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
S24	S6の情報_S13の情報_S7の情報とした文字列	S68	日本語単語対の文字数が同数で編集箇所が両方も漢字の場合、JUMAN辞書の読みが一致するかどうか
S25	S13の情報_S8の情報とした文字列		
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		
S28	S6の情報_S13の情報_S8の情報とした文字列		
S29	S1の字種		
S30	S2の字種		
S31	S3の字種		
S32	S4の字種		
S33	S13の字種		
S34	S14の字種		
S35	S17の字種		
S36	S18の字種		
S37	S1の品詞		
S38	S2の品詞		
S39	S3の品詞		
S40	S4の品詞		
S41	S13の品詞		
S42	S14の品詞		
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:シュウヅン 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 言い替える:言替える 爽かだ:爽快だ ローテーション:ローティション	前:前方 通詞:通辞 極道だ:徹道だ 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 責:責 切り:切間
JUMAN辞書	すじ合:すじ合い よきよう:余きよう なげすてる:投げすてる すり粉木:搦り粉木 けいし:けい復 りょうち:りょう地	洞々たる:洞洞たる やけ残る:焼け残る 糎米:餅米 しがい:市がい 書き出す:書出す 奇れいすぎだ:綺れいすぎだ

【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			

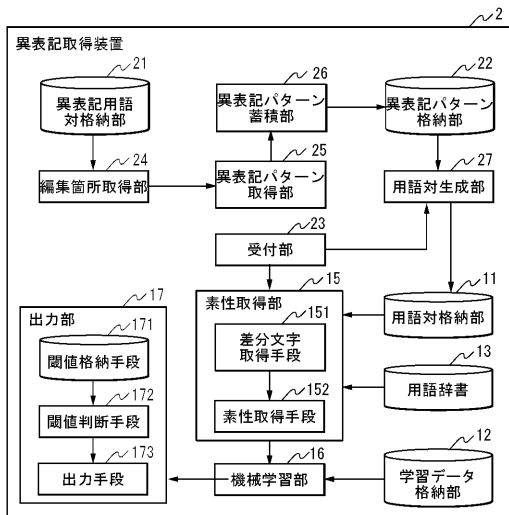
【図17】

	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			

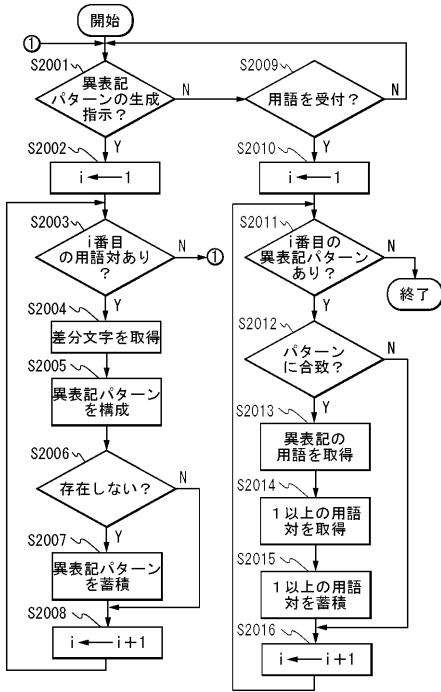
【図18】

	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】



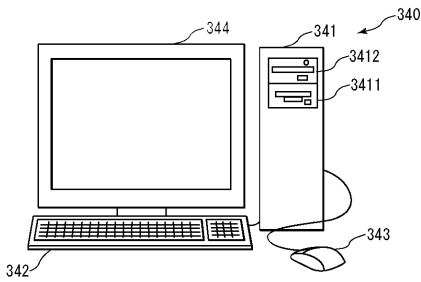
【図20】



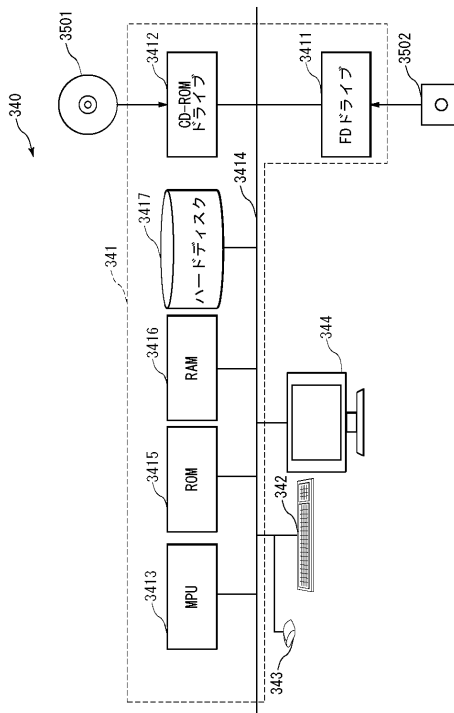
【図21】

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

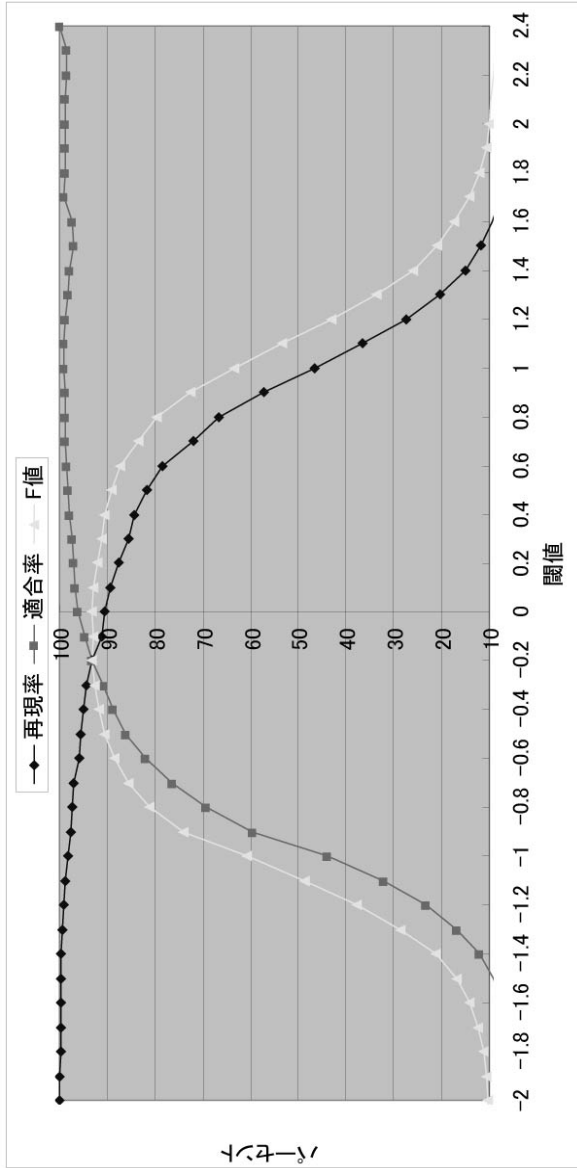
【図22】



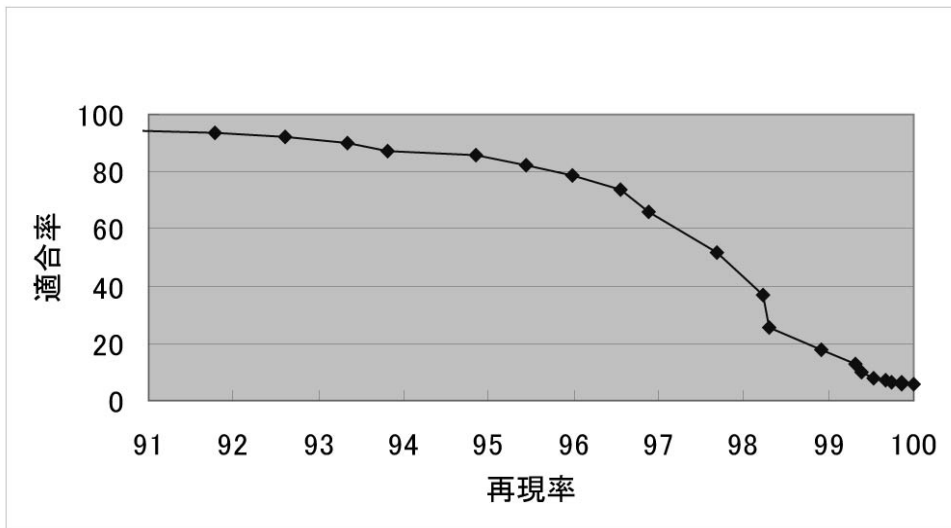
【図23】



【 図 1 4 】



【 図 1 5 】



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

審査官 成瀬 博之

- (56)参考文献 特開2006-004427(JP,A)
特開平03-015980(JP,A)
荒牧英治 他3名, Support Vector Machineを用いた医学用語の表示ゆ
れ解消, 言語処理学会第14回年次大会発表論文集, 日本, 言語処理学会, 2008年 3月1
7日, 135-138頁
佐藤理史, 異表記同語認定のための辞書編纂, 情報処理学会研究報告, 日本, 社団法人情報処理
学会, 2004年 5月13日, Vol.2004, No.47(2004-NL-161), 97-104頁

- (58)調査した分野(Int.Cl., DB名)
G06F 17/20 - 17/30