(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2003－196636

（Ｐ２００３－１９６６３６Ａ）

(43)公開日　平成15年７月11日(2003.7.11)

(51)Int.Cl.[7]	識別記号	ＦＩ			ﾃｰﾏｺｰﾄﾞ (参考)
G06N 3/00	560	G06N 3/00	560	Ｊ	5B009
G06F 17/21	550	G06F 17/21	550	Ｊ	

審査請求　有　請求項の数６　ＯＬ　（全９頁）

(21)出願番号　特願2001－393734(Ｐ2001－393734)

(22)出願日　平成13年12月26日(2001.12.26)

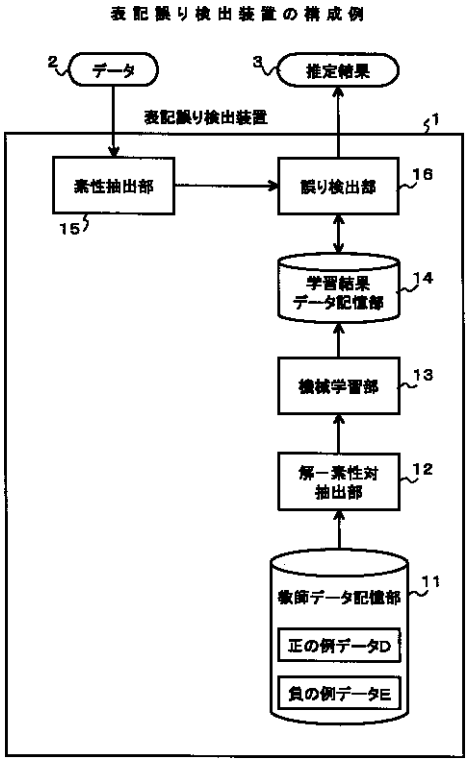
特許法第30条第１項適用申請有り　2001年７月10日　社団法人電子情報通信学会発行の「電子情報通信学会技術研究報告　信学技報　Ｖｏｌ.101　Ｎｏ.190」に発表

(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(74)代理人　100119161
弁理士　重久　啓子　（外１名）

Ｆターム(参考)　5B009 KA05 MG01 QA15

(54)【発明の名称】教師あり機械学習法を用いた表記誤り検出処理方法、その処理装置、およびその処理プログラム

(57)【要約】
【課題】　表記誤り検出処理に関し、教師あり機械学習法を用いて精度の高い検出を行なう。
【解決手段】　教師データとして正の例データＤ（正しい表記）と負の例データＥ（誤りの表記）とを教師データ記憶部11に記憶しておく。解－素性対抽出部12は、教師データ記憶部11の負の例データＥと正の例データＤから解と素性の組を抽出し、機械学習部13は、その素性の集合の場合にどのような解になりやすいかを機械学習法により推定して学習結果データ記憶部14に記憶する。素性抽出部15は、入力したデータ２から素性の集合を抽出し、誤り検出部16は、学習結果データ記憶部１４を参照して、その素性の集合から表記誤りであるかどうかを推定して、推定結果３を出力する。



1

【特許請求の範囲】

【請求項１】　表記の誤りを検出する処理方法であって、

正しい表記である正の例データと誤った表記である負の例データとを含む教師データから素性と解との対を抽出し、抽出した素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理過程と、

入力されたデータから素性を抽出し、前記学習結果データ記憶部に保存された前記学習結果をもとに表記の誤りを検出する処理過程とを備えることを特徴とする教師あり機械学習法を用いた表記誤り検出処理方法。

【請求項２】　表記の誤りを検出する処理方法であって、

入力された事例が予め用意した正しい表記である正の例データに存在しない場合に、前記事例の一般的な出現確率を算出する処理過程と、

前記事例が前記正の例データに出現する確率を前記一般的出現確率をもとに算出し、当該確率が所定のしきい値を超える前記事例を負の例データとする処理過程と、

正の例データと前記負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理過程と、

入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する処理過程とを備えることを特徴とする教師あり機械学習法を用いた表記誤り検出処理方法。

【請求項３】　表記の誤りを検出する処理装置であって、

正しい表記である正の例データと誤った表記である負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、その学習結果を学習結果データ記憶部に保存する機械学習処理手段と、

入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する誤り検出処理手段とを備えることを特徴とする教師あり機械学習法を用いた表記誤り検出処理装置。

【請求項４】　表記の誤りを検出する処理装置であって、

入力された事例が予め用意した正しい表記である正の例データに存在しない場合に、前記事例の一般的な出現確率を算出する出現確率算出処理手段と、

前記事例が前記正の例データに出現する確率を前記一般的出現確率をもとに算出し、当該確率が所定のしきい値を超える前記事例を負の例データとする負の例取得処理手段と、

正の例データと前記負の例データとを含む教師データから素性と解との対を抽出し、抽出した素性と解との対を

2

教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する機械学習処理手段と、

入力されたデータから素性を抽出し、前記学習結果データ記憶部に保存された前記学習結果をもとに表記の誤りを検出する誤り検出処理手段とを備えることを特徴とする教師あり機械学習法を用いた表記誤り検出処理装置。

【請求項５】　表記の誤りを検出する処理をコンピュータに実行させるためのプログラムであって、

正しい表記である正の例データと誤った表記である負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理と、

入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する処理とを、

コンピュータに実行させることを特徴とする教師あり機械学習法を用いた表記誤り検出処理プログラム。

【請求項６】　表記の誤りを検出する処理をコンピュータに実行させるためのプログラムであって、

入力された事例が予め用意した正しい表記である正の例データに存在しない場合に、前記事例の一般的な出現確率を算出する処理と、

前記事例が前記正の例データに出現する確率を前記一般的出現確率をもとに算出し、当該確率が所定のしきい値を超える前記事例を負の例データとする処理と、

正の例データと前記負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理と、

入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する処理とを、

コンピュータに実行させることを特徴とする教師あり機械学習法を用いた表記誤り検出処理プログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、表記誤り検出処理に関し、特に教師あり機械学習法を用いた表記誤り検出処理方法と、その処理を実現する処理装置と、およびその処理をコンピュータに実行させるためのプログラムとに関する。

【０００２】

【従来の技術】日本語の場合の単語の表記誤り検出は、英語の場合に比べてはるかに難しいものである。英語の場合は単語でわかち書きされているために、基本的に単語辞書と単語末の変形の規則とを用意しておくことにより、ほぼ高精度に単語のスペルチェックを行なうことができる。これに対して、日本語の場合は単語でわかち書きされていないために、単語の表記誤りに限定した処理であっても、高精度に行なうことが困難である。

【０００３】また、表記の誤りとしては、単語表記の誤

りの他に、助詞の「て」「に」「を」「は」の運用誤りなどの文法的な誤りも存在する。

【０００４】日本語の表記誤りの検出の主な従来技術として以下のものがある。

【０００５】単語辞書やひらがな連続を登録した辞書や、連接の条件を記述した辞書にもとづいて表記誤りを検出する従来手法などが、以下の参考文献１～参考文献３に記載されている。これらの従来手法では、単語辞書やひらがな連続を登録した辞書にないものがあらわれると表記誤りであると判定したり、連接の条件を記述した辞書において満足されない連接の出現が存在すると表記誤りであると判定する。

［参考文献１：納富一宏，日本語文書校正支援ツールｈｓｐの開発，情報処理学会　研究発表会（デジタル・ドキュメント），(1997)，pp.9-16 ］

［参考文献２：川原一真　他，コーパスから抽出された辞書を用いた表記誤り検出法，情報処理学会第５４回全国大会，(1997)，pp.2-21-2-22］

［参考文献３：白木伸征　他，大量の平仮名列登録による日本語スペルチェッカの作成、言語処理学会　年次大会，(1997)，pp.445-448］

また、文字単位のｎｇｒａｍを利用した確率モデルにもとづいて各文字列の生起確率を求め、生起確率の低い文字列が出現する箇所を表記誤りと判定する従来手法などが、以下の参考文献４～参考文献６に記載されている。

［参考文献４：荒木哲郎　他，２重マルコフモデルによる日本語文の誤り検出並びに訂正法，情報処理学会自然言語処理研究会　NL97-5，(1997)，pp.29-35］

［参考文献５：松山高明　他，ｎ－ｇｒａｍによるｏｃｒ誤り検出の能力検討のための適合率と再現率の推定に関する実験と考察，言語処理学会　年次大会(1996)，pp.129-132］

［参考文献６：竹内孔一　他，統計的言語モデルを用いたＯＣＲ誤り修正システムの構築，情報処理学会論文誌，Vol.40，No.6，(1999)］

上記の従来手法のうち、参考文献５のｎｇｒａｍ確率を利用する手法は、主に光学式文字読み取り装置（Optical Character Reader：ＯＣＲ）の誤り訂正システムにおける表記誤り検出に用いられているものである。ＯＣＲ誤り訂正システムの場合は、前提として表記誤りの出現率が５～１０％と高く、普通に人がものを書くときに誤る確率より高い。したがって、表記誤りの検出の再現率、適合率は高くなりやすく、比較的容易な問題の設定となる。

【０００６】また、上記の従来手法の中で最も良さそうに思われる竹内らの方法、すなわち参考文献６に記載されている従来手法（以下、従来手法Ａという。）を、以下で簡単に説明する。

【０００７】従来手法Ａでは、まず、表記誤りを検出したいテキストを頭から一文字ずつずらしながら3 文字連続を抽出し、抽出した部分のコーパス(正しい日本語文の集合）での出現確率がＴｐ以下の場合に、その各３文字連続に－１を加えていき、与えられた値がＴｓ以上となった文字を誤りと判定する。例えば、Ｔｐ＝０、Ｔｓ＝－２とする。ここで、Ｔｐ＝０としているために、出現確率をわざわざ求める必要はなく、コーパスにその３文字連続が出現するか否かを調べるということをするだけでよい。Ｔｐ＞０とした場合は、抽出した部分がコーパスに出現するものがあっても誤りと判定するものとなる。しかし、出現確率が低くともコーパスに出現していれば、それは誤りとしなくてよいだろうからＴｐ＞０は適切ではなく、Ｔｐ＝０の設定は良いとする。

【０００８】従来手法Ａの補足説明として、「負の事零の検出」という日本語表現に対して誤り検出を行なうことを考える。このとき、日本語表現の頭から「負の事」「の事零」といった連続する３文字を切り出し、これらがコーパスにあるかどうかを調べ、切り出した３文字がなければその３文字に－１を与える。この場合「の事零」「事零の」がなかったため、図７に示すようなｔｒｉｇｒａｍによる得点が与えられ、結果として「－２」点となった「事」と「零」の部分が誤りと判定される。この従来手法Ａは、コーパスに高頻度に出現する文字３－ｇｒａｍをうまく組み合わせて誤りを検出する方法となっている。

【０００９】しかし、結局のところ、従来手法Ａの処理は、コーパスにその表現が存在するか否かを判定するものである。すなわち、従来手法Ａは、辞書にないものがあらわれると誤りとする上記の他の従来手法とよく似たものである。

【００１０】機械学習法については、以下の参考文献７に述べられているように、正の例のみからの学習は一般的に困難であることが知られている。

［参考文献７：横森貫　他，型式言語の学習－正の例からの学習を中心に－，情報処理学会誌，Vol.32，No.3，(1991)，pp226-235 ］

さらに、教師信号とする誤った表記データ（負の例）は、正しい表記データ（正の例）に比べて一般的に取得することが困難であると考えられている。

【００１１】

【発明が解決しようとする課題】従来は、正の例のみを教師信号とする機械学習法を用いた処理方法では高い精度の処理が期待できないこと、および、教師信号とする負の例の取得が困難であることから、文章の表記誤り検出処理において、正の例および負の例の両方を教師信号とした機械学習法を利用した処理方法は実現されていなかった。

【００１２】本発明の目的は、正の例および負の例を教師信号とする機械学習法を用いて、精度の高い表記誤り検出処理を実現することである。

【００１３】また、本発明の別の目的は、教師信号とす

5

る負の例を効率よく自動生成し、機械学習法の教師信号として用いて、精度の高い表記誤り検出処理を実現することである。

【００１４】

【課題を解決するための手段】上記の課題を解決するため、本発明は、表記の誤りを検出する処理方法であって、正しい表記である正の例データと誤った表記である負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理過程と、入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する処理過程とを備える。

【００１５】また、本発明は、表記の誤りを検出する処理方法であって、入力された事例が予め用意した正しい表記である正の例データに存在しない場合に、前記事例の一般的な出現確率を算出する処理過程と、前記事例が前記正の例データに出現する確率を前記一般的出現確率をもとに算出し、当該確率が所定のしきい値を超える前記事例を負の例データとする処理過程と、正の例データと前記負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理過程と、入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する処理過程とを備える。

【００１６】さらに、本発明は、表記の誤りを検出する処理装置であって、正しい表記である正の例データと誤った表記である負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する機械学習処理手段と、入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する誤り検出処理手段とを備える。

【００１７】また、本発明は、表記の誤りを検出する処理装置であって、入力された事例が予め用意した正しい表記である正の例データに存在しない場合に、前記事例の一般的な出現確率を算出する出現確率算出処理手段と、前記事例が前記正の例データに出現する確率を前記一般的出現確率をもとに算出し、当該確率が所定のしきい値を超える前記事例を負の例データとする負の例取得処理手段と、正の例データと前記負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する機械学習処理手段と、入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する誤り検出処理手段とを備える。

【００１８】さらに、本発明は、表記の誤りを検出する処理をコンピュータに実行させるためのプログラムであって、正しい表記である正の例データと誤った表記である負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理と、入力されたデータから素性を抽出し、前記学習結果をもとに前記データの表記の誤りを検出する処理とを、コンピュータに実行させるものである。

【００１９】また、本発明は、表記の誤りを検出する処理をコンピュータに実行させるためのプログラムであって、入力された事例が予め用意した正しい表記である正の例データに存在しない場合に、前記事例の一般的な出現確率を算出する処理と、前記事例が前記正の例データに出現する確率を前記一般的出現確率をもとに算出し、当該確率が所定のしきい値を超える前記事例を負の例データとする処理と、正の例データと前記負の例データとを含む教師データから素性と解との対を抽出し、前記素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データ記憶部に保存する処理と、入力されたデータから素性を抽出し、前記学習結果データ記憶部に保存された前記学習結果をもとに前記データの表記の誤りを検出する処理とを、コンピュータに実行させるものである。

【００２０】本発明の各手段または機能または要素をコンピュータにより実現するプログラムは、コンピュータが読み取り可能な、可搬媒体メモリ、半導体メモリ、ハードディスクなどの適当な記録媒体に格納することができ、これらの記録媒体に記録して提供され、または、通信インタフェースを介して種々の通信網を利用した送受信により提供される。

【００２１】

【発明の実施の形態】以下に、本発明の第１の実施の形態として、日本語表記誤りを連接により検出する処理を説明する。

【００２２】図１に、本発明の第１の実施の形態における表記誤り検出装置１の構成例を示す。

【００２３】表記誤り検出装置１は、教師データ記憶部１１と、解－素性対抽出部１２と、機械学習部１３と、学習結果データ記憶部１４と、素性抽出部１５と、誤り検出部１６とを持つ。

【００２４】教師データ記憶部１１は、機械学習法を実施する際の教師信号となるデータ（教師データ）を記憶する手段である。教師データ記憶部１１には、教師データとして、正しい表記である事例（正の例）と誤った表記である事例（負の例）とが記憶される。正の例は、例えば正しい文の集合であるコーパス等を利用してもよい。負の例は、誤った表記であって一般的なデータはないため、予め人手により生成したものを用いる。または、後述するような負の例予測処理方法を用いて正の例から生成するようにしてもよい。

【００２５】解－素性対抽出部１２は、教師データ記憶

部１１に記憶されている教師データの各事例ごとに、事例の解と素性の集合との組を抽出する手段である。

【００２６】機械学習部１３は、解－素性対抽出部１２により抽出された解と素性の集合の組から、どのような素性のときにどのような解になりやすいかを機械学習法により学習する手段である。その学習結果は、学習結果データ記憶部１４に保存される。

【００２７】素性抽出部１５は、表記誤り検出対象であるデータ２から素性の集合を抽出し、抽出した素性の集合を誤り検出部１６へ渡す手段である。

【００２８】誤り検出部１６は、学習結果データ記憶部１４の学習結果データを参照して、素性抽出部１５から渡された素性の集合の場合に、どのような解になりやすいか、すなわち表記誤りであるかどうかを推定し、その推定結果３を出力する手段である。

【００２９】図２に、教師データ記憶部１１のデータ構成例を示す。教師データ記憶部１１には、問題と解との組である教師データが記憶されている。例えば、文の各文字のすき間（＜｜＞で表す。）を問題として、そのすき間の連接の解（正解、誤り）が対応付けられた教師データが記憶される。図２の教師データのうち、

「問題－解：説明した方法で＜｜＞を用いることができる－誤り」

は、負の例データＥの例であり、

「問題－解：説明した方法＜｜＞でを用いることができる－正」

は、正の例データＤの例である。

【００３０】図３に、表記誤り検出処理の処理フローチャートを示す。表記誤り検出処理前に、正の例データＤおよび負の例データＥが教師データ記憶部１１に記憶されているとする。

【００３１】まず、解－素性対抽出部１２は、教師データ記憶部１１から、各事例ごとに、解と素性の集合との組を抽出する（ステップＳ１）。素性とは、解析に用いる情報の細かい１単位を意味する。素性として連接の判定対象となる文字のすき間ごとに以下のものを抽出する。

【００３２】・前接および後接の各１～５ｇｒａｍの文字列、

・対象（すき間）を含めた１～５ｇｒａｍの文字列（ただし、対象であるすき間（＜｜＞）は１文字として扱う。）

・前接および後接の単語（単語の抽出は既存の形態素解析処理を行う処理手段（図１には図示しない）などを利用する。）

・前接および後接の単語の品詞

例えば、「問題－解」が、「説明した方法で＜｜＞を用いることができる－誤り」である場合には、図４に示すような素性を抽出する。すなわち、以下の素性を抽出する。

【００３３】素性：前接「した方法で」，前接「た方法で」，前接「方法で」，前接「法で」，前接「で」，後接「を用いるこ」，後接「を用いる」，後接「を用い」，後接「を用」，後接「を」，「た方法で＜｜＞」，「方法で＜｜＞を」，「法で＜｜＞を用」，「で＜｜＞を用い」，「＜｜＞を用いる」，前接「で」，後接「を」，前接「助詞」，後接「助詞」

次に、機械学習部１３は、抽出した解と素性の集合との組から、どのような素性のときにどのような解になりやすいかを機械学習し、その学習結果を学習結果データ記憶部１４に保存する（ステップＳ２）。

【００３４】機械学習の手法としては、例えば、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を用いる。

【００３５】決定リスト法は、素性と分類先の組を規則とし、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。

【００３６】最大エントロピー法は、あらかじめ設定しておいた素性ｆｊ（１≦ｊ≦ｋ）の集合をＦとするとき、所定の条件式を満足しながらエントロピーを意味する式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、もっとも大きい確率値を持つ分類を求める分類とする方法である。

【００３７】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、２つの分類からなるデータを分類する手法である。

【００３８】決定リスト法および最大エントロピー法については、以下の参考文献８に、サポートベクトルマシン法については、以下の参考文献９および参考文献１０に説明されている。

［参考文献８：村田真樹、内山将夫、内元清貴、馬青、井佐原均、種々の機械学習法を用いた多義解消実験、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，NCL2001-2，(2001)］

［参考文献９：Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and other kernel-based learning methods,(Cambridge University Press,2000)］

［参考文献１０：Taku Kudoh, Tinysvm:Support Vector machines,(http://cl.aist-nara.ac.jp/taku-ku//software/Tiny SVM/index.html,2000)］

なお、機械学習部１３では、上記の手法に限定されずに、教師あり機械学習法であればどのような手法でも使用することができる。

【００３９】その後、素性抽出部１５は、解を求めたいデータ２を入力し（ステップＳ３）、解－素性対抽出部

１２での処理とほぼ同様に、データ２から素性の集合を取り出し、それらを誤り検出部１６へ渡す（ステップＳ４）。

【００４０】誤り検出部１６は、渡された素性の集合の場合にどのような解になりやすいかを学習結果データ記憶部１４の学習結果データをもとに特定し、特定した解すなわち表記誤りかどうかの推定結果３を出力する（ステップＳ５）。

【００４１】例えば、解析したい問題がすき間＜｜＞の連接である場合に、データ２が「説明した方法で＜｜＞を用いることができる」であれば、「誤り」という推定結果３を出力する。

【００４２】次に、本発明の第２の実施の形態について説明する。

【００４３】教師データ記憶部１１の正の例データＤについては、コーパス等を利用できるため比較的容易に取得できる。しかし、負の例データＥは、容易に取得できないため人手により生成するが、かかる生成作業の負担は大きい。

【００４４】また、教師データは多量であるほうが処理精度が向上するため、できる限り多量の教師データを用意することが望ましい。

【００４５】そこで、多量な正の例データから負の例データを予測する方法を考える。

【００４６】正の例から負の例を予測する単純な方法として、既知の正の例のデータに現れなかったものをすべて負の例とするという手法が考えられる。しかし、実際には未出現の正の例の存在が考えられるために、このような単純な方法を用いると、多くの未出現の正の例を負の例であると判定してしまうことになるという問題があり、このような方法で生成した負の例を高精度の処理に適用することができない。

【００４７】例えば大規模な既存のコーパス（日本語の文の集合）をすべて正しいと仮定すると、その既存のコーパスを正しい文（正の例）と考え、この正の例を用いて、表記誤り（負の例）を予測する方法により、自動的に負の例を生成することができる。

【００４８】これにより、教師データとする負の例が豊富になり、生成作業の負担を軽減し、かつ、教師データ付きの機械学習法を利用した高精度の表記誤り検出処理を実現できることになる。

【００４９】本形態における表記誤り検出装置１は、まず、正の例か負の例か判定すべき未知の事例ｘの一般的な出現確率ｐ（ｘ）を算出する。次に、この出現確率ｐ（ｘ）で既知の正の例データＤに出現しないことが不自然である場合に、すなわち、一般的な出現確率が高く当然正の例データＤに出現するであろう状態にも関わらず既知の正の例データＤに出現しない場合には、事例ｘの負の例の度合いが高いと推測し、所定の値より高い負の例の度合いの事例ｘを負の例データＥとする。そして、かかる負の例データＥと正の例データＤとを教師信号とした機械学習法により表記誤り検出処理を行う。

【００５０】図５に、本発明の第２の実施の形態における表記誤り検出装置１の構成例を示す。

【００５１】表記誤り検出装置１は、教師データ記憶部１１と、解－素性対抽出部１２と、機械学習部１３と、素性抽出部１５と、誤り検出部１６と、存在判定部２１と、出現確率推定部２２と、負の例度合い算出部２３と、負の例取得部２４と、正の例データ記憶部２５とを持つ。

【００５２】教師データ記憶部１１と、解－素性対抽出部１２と、機械学習部１３と、素性抽出部１５と、誤り検出部１６とは、第１の実施の形態で説明した表記誤り検出装置１の各手段と同一の手段であるので説明を省略する（図１参照）。

【００５３】存在判定部２１は、正または負の情報が付与されていない日本語文の集合であるコーパス２０の事例ｘが、正の例データ記憶部２５に記憶されている正の例データＤに存在するかどうかを判定する手段である。

【００５４】出現確率推定部２２は、事例ｘが正の例データ記憶部２５に存在しない場合に、事例ｘの一般的な出現確率（頻度）ｐ（ｘ）を算出する手段である。

【００５５】負の例度合い算出部２３は、出現確率ｐ（ｘ）をもとに事例ｘの負の例度合いＱ（ｘ）を算出する手段である。

【００５６】負の例取得部２４は、負の例度合い算出部２３から受け取った事例ｘの負の例度合いＱ（ｘ）が所定の値を超える場合に、その事例ｘを負の例データＥとし、事例ｘを問題－解の構想の教師データ（負の例データＥ）として教師データ記憶部１１に記憶する手段である。

【００５７】図６に、第２の実施の形態において学習データとなる負の例データの取得処理の処理フローチャートを示す。

【００５８】表記誤り検出装置１の存在判定部２１は、コーパス２０から正の例か負の例かが未知である文を入力し、文の頭から、文字のすき間を１つずつずらしながら、各すき間を連接チェックの対象として、そのすき間に前接する１～５ｇｒａｍの文字列ａと、後接する１～５ｇｒａｍの文字列ｂを取り出し、この任意のペアである事例ｘ＝（ａ、ｂ）を生成する（ステップＳ１１）。ここでは、２５個の事例（ペア）が生成されることになる。

【００５９】そして、事例ｘの２５個の連接ａｂが正の例データ記憶部２５にあるかどうかを調べ（ステップＳ１２）、連接ａｂが正の例データ記憶部２５に存在しなければ、その事例ｘを出現確率推定部２２へ渡す（ステップＳ１３）。

【００６０】出現確率推定部２２は、事例ｘの一般的な出現確率ｐ（ｘ）を推定する（ステップＳ１４）。

【００６１】例えば、正の例データ記憶部２５の正の例データＤは二項関係（ａ，ｂ）からなり、二項のａとｂとがお互いに独立であると仮定すると、二項関係（ａ，ｂ）の出現する確率はｐ（ｘ）は、ａ、ｂの正の例データ記憶部２５での出現確率をｐ（ａ）、ｐ（ｂ）とするとき、その積ｐ（ａ）×ｐ（ｂ）となる。すなわち、各事例ｘを二項関係（ａ，ｂ）とし、その各項ａ、ｂを独立と仮定することで、各事例ｘの一般的な出現確率ｐ（ｘ）を、各項ａ、ｂの確率により計算する。

【００６２】そして、負の例度合い算出部２３は、事例ｘの出現確率ｐ（ｘ）を使って、事例ｘが正の例データ記憶部２５に出現する確率Ｑ（ｘ）を求める（ステップＳ１５）。

【００６３】このとき、正の例データ記憶部２５の正の例データＤがｎ個でありそれぞれが独立であることを仮定すると、１回試行して事例ｘが出現しない確率は１－ｐ（ｘ）であり、これがｎ回連続して起こるということから、事例ｘが正の例データ記憶部２５の正の例データＤに出現しない確率は（１－ｐ（ｘ））^{n} となり、事例ｘが同じく正の例データＤに出現する確率Ｑ（ｘ）＝１－（１－ｐ（ｘ））^{n} となる。

【００６４】ところで、「確率Ｑ（ｘ）が小さい」というのは、確率的に事例ｘが正の例データ記憶部２５の正の例データＤに出現する確率が低いということであり、正の例データ（コーパス）が小さいために確率的に出現しないということが保証されたことを意味するため、「事例ｘは正の例でありうる。」という意味になる。

【００６５】逆に、「確率Ｑ（ｘ）が大きい」というのは、確率的に事例ｘが正の例データＤに出現する確率が高いということであり、確率的には同コーパスに当然出現すべきということになり、それなのに実際は出現しなかったということで矛盾が生じることになる。この矛盾により、一般的な出現確率ｐ（ｘ）か種々の独立の仮定が否定されることになる。

【００６６】ここで、「事例ｘが正の例である場合は、一般的な出現確率ｐ（ｘ）および種々の独立の仮定が正しい。」と新たに仮定すると、この矛盾により「事例ｘは正の例でありえない。」が導出されることになる。すなわち、「事例ｘが正の例データＤに出現する確率Ｑ（ｘ）」は、「事例ｘが正の例でありえない確率Ｑ（ｘ）」を意味することになる。そういう意味で、Ｑ（ｘ）は負の例の度合いを意味するものとなる。よって、このＱ（ｘ）を「負の例度合い」とし、事例ｘのＱ（ｘ）が大きいほど事例ｘの負の例の度合いが大きいとする。

【００６７】そして、負の例取得部２４は、最もＱ（ｘ）の値が高いときのその値をQ_{max}、またｘをx_{max} とし、Ｑ（x_{max} ）の値が大きいすき間ほど、妥当でない連接の可能性が高いとして、Ｑ（x_{max} ）の値が、所定の値より大きい場合には、そのすき間を負の例データＥとして教師データ記憶部１１へ保存する（ステップＳ１６）。なお、負の例データＥとその負の例の度合いＱ（x_{max} ）とを教師データ記憶部１１に保存してもよい。

【００６８】以上のステップＳ１１～ステップＳ１５の処理を、文の全てのすき間について行っていくことにより、正の例データ記憶部２５の正の例データＤの頻度情報を用いて負の例データＥを取得することができ、正の例データＤおよび負の例データＥを教師データとして教師データ記憶部１１に用意することができる。

【００６９】以降の処理は、第１の実施の形態で説明した誤り検出処理と同様であるので、説明を省略する。

【００７０】以上、本発明をその実施の形態により説明したが、本発明はその主旨の範囲において種々の変形が可能である。

【００７１】例えば、表記誤り検出装置１の出現確率推定部２２は、事例ｘの一般的な出現確率ｐ（ｘ）を、何らかの方法で算出すればよく、本発明の実施の形態で説明した方法に限られるものではない。

【００７２】また、教師データ記憶部１１の正の例データＤは、正の例データ記憶部２５に記憶されている正の例データＤを使用することもでき、また、別に用意した正の例データを使用することもできる。

【００７３】

【発明の効果】以上説明したように、本発明は、正の例と負の例とを教師信号とする機械学習法を用いて表記誤り検出処理を行う。負の例の情報も用いる本発明は、正の例だけを用いた処理方法に比べて、格段に高い精度の処理結果を得ることができる。

【００７４】また、本発明は、正の例の頻度情報を用いて、正の例から負の例を抽出する処理を行い、抽出した負の例を機械学習法の教師信号とする。正の例から自動的に抽出される負の例の情報を用いる本発明は、表記誤り検出のように正の例が存在するが負の例の取得が困難な問題において、負の例を生成する処理負担を軽減することができる。

【図面の簡単な説明】

【図１】第１の実施の形態における表記誤り検出装置の構成例を示す図である。

【図２】教師データ記憶部の構成例を示す図である。

【図３】表記誤り検出処理の処理フローチャート図である。

【図４】素性の例を示す図である。

【図５】第２の実施の形態における表記誤り検出装置の構成例を示す図である。

【図６】負の例データ取得処理の処理フローチャート図である。

【図７】従来手法を補足的に説明するための図である。

【符号の説明】

１　表記誤り検出装置

13

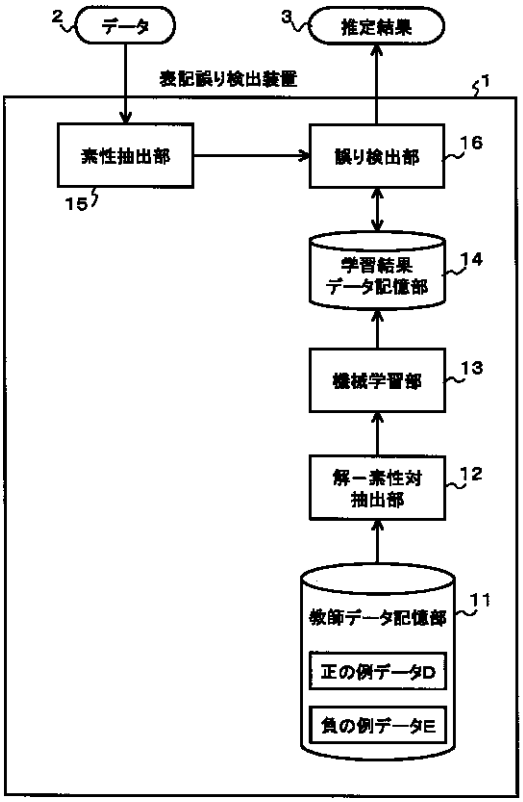
２　データ
３　推定結果
１１　教師データ記憶部
１２　解－素性対抽出部
１３　機械学習部
１４　学習結果データ記憶部
１５　素性抽出部

14

１６　誤り検出部
２０　コーパス
２１　存在判定部
２２　出現確率推定部
２３　負の例度合い算出部
２４　負の例取得部
２５　正の例データ記憶部

【図１】

表記誤り検出装置の構成例



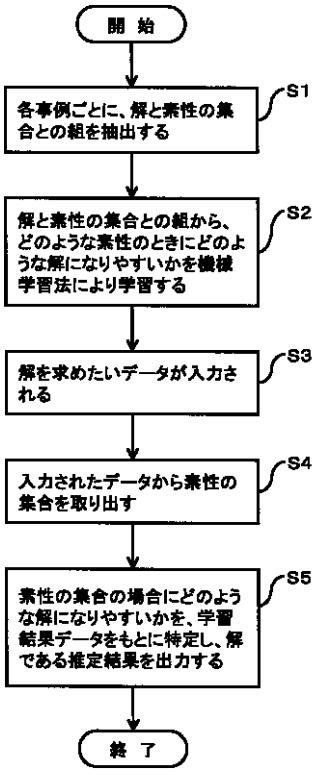
【図２】

教師データ記憶部の構成例

問　題	解
説明した方法で＜｜＞を用いることができる	誤り
説明した方法＜｜＞でを用いることができる	正
⋮	⋮

＜｜＞:すき間

【図３】



【図４】

素性の例

前接「した方法で」
前接「た方法で」
前接「方法で」
前接「法で」
前接「で」
後接「を用いるこ」
後接「を用いる」
後接「を用い」
後接「を用」
後接「を」
「た方法で＜｜＞」
「方法で＜｜＞を」
「法で＜｜＞を用」
「で＜｜＞を用い」
「＜｜＞を用いる」
前接「で」
後接「を」
前接「助詞」
後接「助詞」

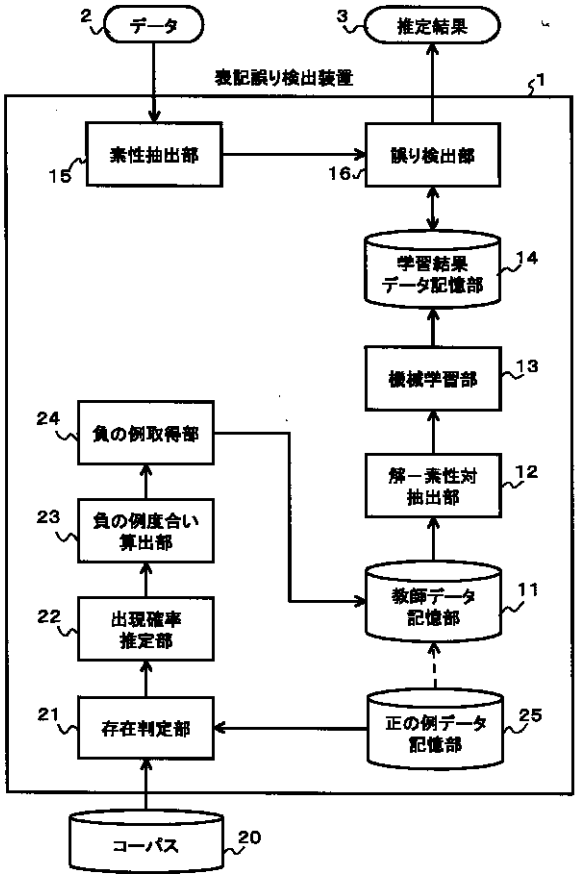
【図７】

例　文	負　の　事　例　の　検　出
trigramによる得点	－１　－１　－１ －１　－１　－１
合計得点	－１　－２　－２　－１

誤り

【図５】

表記誤り検出装置の構成例



【図６】

負の例データ取得処理の処理フローチャート