

(51)Int.Cl. ⁷	識別記号	F I	テ-マコード [*]	(参考)
G06N 3/00	560	G06N 3/00	J 5B009	
G06F 17/21	550	G06F 17/21	J	

審査請求 有 請求項の数 6 O L (全9頁)

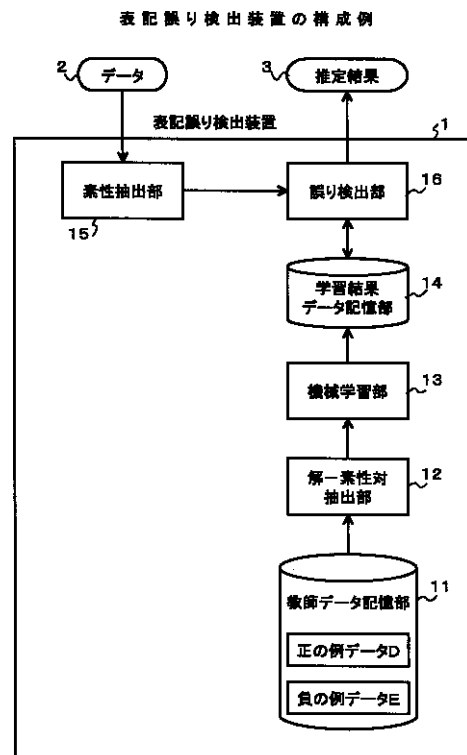
(21)出願番号	特願2001 - 393734(P 2001 - 393734)	(71)出願人	301022471 独立行政法人通信総合研究所 東京都小金井市貫井北町4 - 2 - 1
(22)出願日	平成13年12月26日(2001.12.26)	(72)発明者	村田 真樹 東京都小金井市貫井北町4 - 2 - 1 独立 行政法人通信総合研究所内
特許法第30条第1項適用申請有り 2001年7月10日 社 団法人電子情報通信学会発行の「電子情報通信学会技術 研究報告 信学技報 V o l .101 N o .190」に発表		(72)発明者	井佐原 均 東京都小金井市貫井北町4 - 2 - 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。



【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つて、
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つて、

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な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つて、

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つて、

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な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

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つて、

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つて、

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な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

10

20

30

40

50

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、接続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、接続の条件を記述した辞書において満足されない接続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クの作成，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ることで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

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な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な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な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、日本語表記誤りを接続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、そのすき間の接続の解(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。素性として接続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」、「<|>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された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 $f_j(1 \leq j \leq k)$ 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< | >の接続である場合に、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、各すき間を接続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個の接続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)、接続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 $p(a) \times p(b)$ 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 $p(x)$ が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、妥当でない接続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

- 2 データ
- 3 推定結果
- 1 1 教師データ記憶部
- 1 2 解 - 素性対抽出部
- 1 3 機械学習部
- 1 4 学習結果データ記憶部
- 1 5 素性抽出部

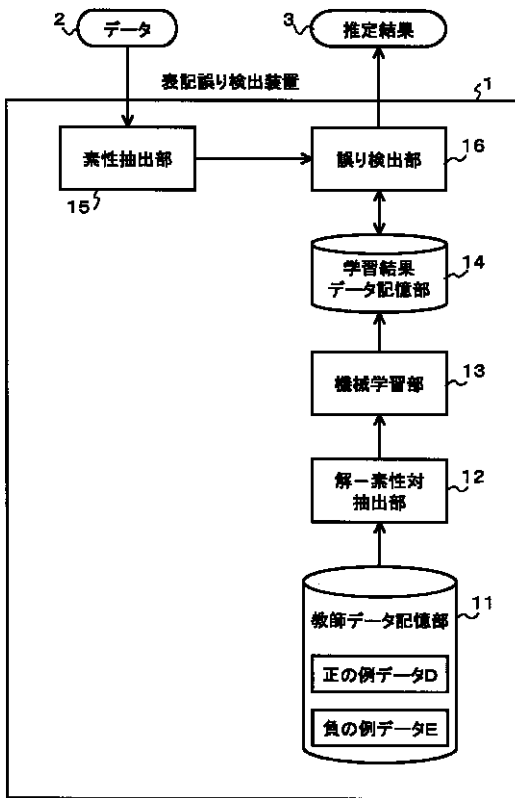
- 1 6 誤り検出部
- 2 0 コーパス
- 2 1 存在判定部
- 2 2 出現確率推定部
- 2 3 負の例度合い算出部
- 2 4 負の例取得部
- 2 5 正の例データ記憶部

【図 1】

【図 2】

【図 4】

表記誤り検出装置の構成例



教師データ記憶部の構成例

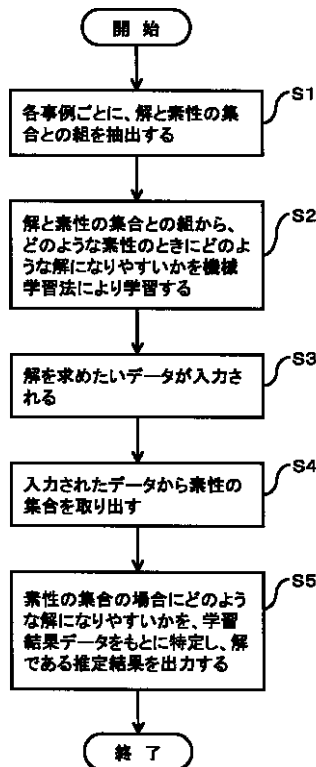
問題	解
説明した方法で< >を用いることができる	誤り
説明した方法< >でを用いることができる	正
⋮	⋮

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」

<|>:すき間

【図 3】

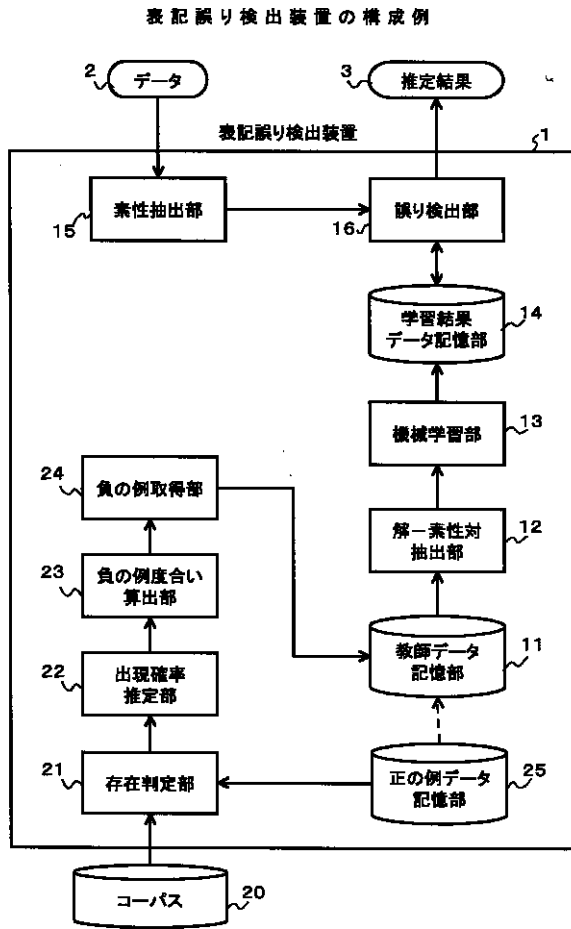


【図 7】

例文	負の事象の検出
trigramによる得点	-1 -1 -1 -1 -1 -1
合計得点	-1 -2 -2 -1

↑
誤り

【 図 5 】



【 図 6 】

負の例データ取得処理の処理フローチャート

