

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第3783053号
(P3783053)

(45) 発行日 平成18年6月7日(2006.6.7)

(24) 登録日 平成18年3月24日(2006.3.24)

(51) Int. Cl.	F I		
G06N 5/04 (2006.01)	G06N 5/04	550N	
G06F 17/21 (2006.01)	G06F 17/21	550A	
G06F 17/27 (2006.01)	G06F 17/27	X	

請求項の数 16 (全 21 頁)

<p>(21) 出願番号 特願2001-394112 (P2001-394112)</p> <p>(22) 出願日 平成13年12月26日(2001.12.26)</p> <p>(65) 公開番号 特開2003-196094 (P2003-196094A)</p> <p>(43) 公開日 平成15年7月11日(2003.7.11)</p> <p>審査請求日 平成13年12月26日(2001.12.26)</p> <p>特許法第30条第1項適用 平成13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</p>	<p>(73) 特許権者 301022471 独立行政法人情報通信研究機構 東京都小金井市貫井北町4-2-1</p> <p>(74) 代理人 100119161 弁理士 重久 啓子</p> <p>(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</p> <p>(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</p> <p>審査官 赤穂 州一郎</p> <p style="text-align: right;">最終頁に続く</p>
---	---

(54) 【発明の名称】 負の例予測処理方法、処理プログラムおよび処理装置、負の例予測処理を用いた日本語表記誤り検出処理プログラムおよび処理装置、ならびに負の例予測処理を用いた他の関係検出処理プログラム

(57) 【特許請求の範囲】

【請求項1】

ある問題について正または負であることが未知のデータについて負であるか否かを予測する処理を行うために、前記問題について正の例である正の例データ群を記憶する正の例データ記憶手段と、データ入力処理手段、存在判定処理手段、出現確率算出処理手段、および負の例度合い算出処理手段とを備えたコンピュータが行う処理方法であって、

前記データ入力処理手段が、予測処理の対象となる2つの項で構成されるデータを取得するデータ入力処理過程と、

前記存在判定処理手段が、前記正の例データ記憶手段に格納された正の例データ群を参照して、前記データ入力処理過程において取得された入力データが前記正の例データ群に存在するか否かを判定する存在判定処理過程と、

前記出現確率算出処理手段が、前記入力データが前記正の例データ記憶手段に格納された正の例データ群に存在しない場合に、前記入力データの各項が前記正の例データ記憶手段に格納された正の例データ群に出現する出現確率を算出し、当該入力データの各項の出現確率の積にもとづいて前記入力データが前記正の例データ群に出現する出現確率を算出する出現確率算出処理過程と、

前記負の例度合い算出処理手段が、前記出現確率算出処理過程において算出された入力データの出現確率をもとに、前記入力データが前記問題について負の例である傾向を示す負の例度合いを求める負の例度合い算出処理過程とを備える

ことを特徴とする負の例予測処理方法。

10

20

【請求項 2】

前記コンピュータは、負の例出力処理手段を備え、
前記負の例出力処理手段が、前記入力データが前記正の例データ記憶手段に格納された正の例データ群に存在しない場合に、前記入力データおよびその負の例度合いを出力する負の例出力処理過程を備える

ことを特徴とする請求項 1 記載の負の例予測処理方法。

【請求項 3】

前記負の例出力処理過程において、前記負の例出力処理手段が、前記出力する入力データが複数である場合に、前記入力データを、その負の例度合いにもとづいて降順もしくは昇順に並べ替えて出力する処理を行う

ことを特徴とする請求項 2 記載の負の例予測処理方法。

【請求項 4】

前記コンピュータは、負の例表示処理手段を備え、
前記負の例表示処理手段が、前記出力する入力データを、その負の例度合いに対応する所定の区分けに応じて、色もしくは輝度を変更して表示し、または、異なる表示形態によって表示する負の例表示処理過程を備える

ことを特徴とする請求項 1 記載の負の例予測処理方法。

【請求項 5】

ある問題について正または負であることが未知のデータについて負であるか否かを予測する処理を、前記問題について正の例である正の例データ群を記憶する正の例データ記憶手段を備えるコンピュータに実行させるためのプログラムであって、

前記正の例データ記憶手段にアクセスする処理と、

予測処理の対象となる 2 つの項で構成されるデータを取得するデータ入力処理と、

前記正の例データ記憶手段に格納された正の例データ群を参照して、前記データ入力処理によって取得された入力データが前記正の例データ群に存在するか否かを判定する存在判定処理と、

前記入力データが前記正の例データ記憶手段に格納された正の例データ群に存在しない場合に、前記入力データの各項が前記正の例データ記憶手段に格納された正の例データ群に出現する出現確率を算出し、当該入力データの各項の出現確率の積にもとづいて前記入力データが前記正の例データ群に出現する出現確率を算出する出現確率算出処理と、

前記出現確率算出処理によって算出された入力データの出現確率をもとに、前記入力データが前記問題について負の例である傾向を示す負の例度合いを求める処理とを、

前記コンピュータに実行させるための負の例予測処理プログラム。

【請求項 6】

前記入力データが前記正の例データ記憶手段に格納された正の例データ群に存在しない場合に、前記入力データおよびその負の例度合いを出力する処理を、

前記コンピュータに実行させるための請求項 5 記載の負の例予測処理プログラム。

【請求項 7】

前記負の例出力処理において、前記出力する入力データが複数である場合に、前記入力データを、その負の例度合いにもとづいて降順もしくは昇順に並べ替えて出力する処理を、

前記コンピュータに実行させるための請求項 6 記載の負の例予測処理プログラム。

【請求項 8】

前記出力する入力データを、その負の例度合いに対応する所定の区分けに応じて、色もしくは輝度を変更して表示し、または、異なる表示形態によって表示する負の例表示処理を、

前記コンピュータに実行させるための請求項 5 記載の負の例予測処理プログラム。

【請求項 9】

ある問題について正または負であることが未知のデータについて負であるか否かを予測する処理を行うコンピュータである負の例予測処理装置であって、

10

20

30

40

50

前記問題について正の例である正の例データ群を記憶する正の例データ記憶手段と、
 予測処理の対象となる2つの項で構成されるデータを取得するデータ入力処理手段と、
 前記正の例データ記憶手段に格納された正の例データ群を参照して、前記データ入力処理手段によって取得された入力データが前記正の例データ群に存在するか否かを判定する存在判定処理手段と、

前記入力データが前記正の例データ記憶手段に格納された正の例データ群に存在しない場合に、前記入力データの各項が前記正の例データ記憶手段に格納された正の例データ群に出現する出現確率を算出し、当該入力データの各項の出現確率の積にもとづいて前記入力データが前記正の例データ群に出現する出現確率を算出する出現確率算出処理手段と、

前記出現確率算出処理手段によって算出された入力データの出現確率をもとに、前記入力データが前記問題について負の例である傾向を示す負の例度合いを求める負の例度合い算出処理手段とを備える

ことを特徴とする負の例予測処理装置。

【請求項10】

前記入力データが前記正の例データ記憶手段に格納された正の例データ群に存在しない場合に、前記入力データおよびその負の例度合いを出力する負の例出力処理手段を備えることを特徴とする請求項9記載の負の例予測処理装置。

【請求項11】

前記負の例出力処理手段は、前記出力する入力データが複数である場合に、前記入力データを、その負の例度合いにもとづいて降順もしくは昇順に並べ替えて出力する

ことを特徴とする請求項10記載の負の例予測処理装置。

【請求項12】

前記出力する入力データを、その負の例度合いに対応する所定の区分けに応じて、色もしくは輝度を変更して表示し、または、異なる表示形態によって表示する負の例表示処理手段を備える

ことを特徴とする請求項9記載の負の例予測処理装置。

【請求項13】

日本語表記の正誤検出処理において正または負であることが未知のデータについて日本語表記の誤りであることを示す負の例であるか否かを予測する負の例予測処理を用いて、入力されたデータの日本語表記誤りを検出する処理を、正しい日本語表記の例である正の例データ群を記憶する正の例データ記憶手段を備えたコンピュータに実行させるためのプログラムであって、

前記正の例データ記憶手段にアクセスする正の例データアクセス処理と、

検出処理の対象となるデータを取得するデータ入力処理と、

前記データ入力処理によって取得された入力データの文字の各すき間について、前記すき間に接続する前接文字列および後接文字列を項とするペアを生成するペア生成処理と、

前記正の例データ記憶手段に格納された正の例データ群を参照して、前記ペア生成処理によって生成されたペアが前記正の例データ群に存在するか否かを判定する存在判定処理と、

前記入力データから生成されたペアから前記正の例データ記憶手段に格納された正の例データ群に存在しないペアを抽出し、前記抽出されたペアについて、当該ペアの各項が前記正の例データ群に出現する出現確率を算出し、当該ペアの各項の出現確率の積にもとづいて前記ペアが前記正の例データ群に出現する出現確率を算出する出現確率算出処理と、

前記出現確率算出処理によって算出された前記ペアの出現確率をもとに、前記ペアが負の例である傾向を示す負の例度合いを求める負の例度合い算出処理と、

前記負の例度合い算出処理によって求められた前記負の例度合いが所定より大きいペアを表記の誤り箇所として検出する表記誤り検出処理とを、

コンピュータに実行させるための負の例予測処理を用いた日本語表記誤り検出処理プログラム。

【請求項14】

10

20

30

40

50

日本語表記の正誤検出処理において正または負であることが未知のデータについて日本語表記の誤りであることを示す負の例であるか否かを予測する負の例予測処理を用いて日本語表記誤りを検出するコンピュータである処理装置であって、

正しい日本語表記の例である正の例データ群を記憶する正の例データ記憶手段と、

検出処理の対象となるデータを取得し、前記取得された入力データの文字の各すき間について、前記すき間に接続する前接文字列および後接文字列を項とするペアを生成するペア生成処理手段と、

前記正の例データ記憶手段に格納された正の例データ群を参照して、前記ペア生成処理手段によって生成されたペアが前記正の例データ群に存在するか否かを判定する存在判定処理手段と、

10

前記入力データから生成されたペアから前記正の例データ記憶手段に格納された正の例データ群に存在しないペアを抽出し、前記抽出されたペアについて、当該ペアの各項が前記正の例データ群に出現する出現確率を算出し、当該ペアの各項の出現確率の積にもとづいて前記ペアが前記正の例データ群に出現する出現確率を算出する出現確率算出処理手段と、

前記出現確率算出処理手段によって算出された前記ペアの出現確率をもとに、前記ペアが負の例である傾向を示す負の例度合いを求める負の例度合い算出処理手段と、

前記負の例度合い算出処理手段によって求められた前記負の例度合いが所定より大きいペアを表記の誤り箇所として検出する表記誤り検出処理手段とを備える

ことを特徴とする負の例予測処理を用いた日本語表記誤り検出処理装置。

20

【請求項 15】

連体節の格関係が未知のデータについて外の関係を示す負の例であるか否かを予測する負の例予測処理を用いて、入力されたデータの外の関係となる連体節を検出する処理を、格関係にある動詞と名詞との組を持つ正の例データ群を記憶する正の例データ記憶手段を備えたコンピュータに実行させるためのプログラムであって、

前記正の例データ記憶手段にアクセスする正の例データアクセス処理と、

検出処理の対象となるデータを取得するデータ入力処理と、

前記データ入力処理によって取得された入力データの連体修飾関係を構成する連体節の動詞とのかかり先の名詞とを抽出し、前記動詞および前記名詞を項とするペアを生成するペア生成処理と、

30

前記正の例データ記憶手段に格納された正の例データ群を参照して、前記ペア生成処理によって生成されたペアが前記正の例データ群に存在するか否かを判定する存在判定処理と、

前記入力データから生成されたペアから前記正の例データ記憶手段に格納された正の例データ群に存在しないペアを抽出し、前記抽出したペアについて、当該ペアの各項が前記正の例データ群に出現する出現確率を算出し、当該各項の出現確率の積にもとづいて前記ペアが前記正の例データ群に出現する出現確率を算出する出現確率算出処理と、

前記出現確率算出処理によって算出された前記ペアの出現確率をもとに、前記ペアが負の例である傾向を示す負の例度合いを求める負の例度合い算出処理と、

前記負の例度合い算出処理において求められた前記負の例度合いが所定より大きいペアを外の関係の連体節として検出する外の関係検出処理とを、

40

コンピュータに実行させるための負の例予測処理を用いた外の関係検出処理プログラム。

【請求項 16】

連体節の格関係が未知のデータについて外の関係を示す負の例であるか否かを予測する負の例予測処理を用いて、入力されたデータの外の関係となる連体節を検出するコンピュータである処理装置であって、

格関係にある動詞と名詞との組を持つ正の例データ群を記憶する正の例データ記憶処理手段と、

検出処理の対象となるデータを取得し、前記取得された入力データが連体修飾関係を構

50

成する連体節の動詞とのかかり先の名詞とを抽出し、前記動詞および前記名詞を項とするペアを生成するペア生成処理手段と、

前記正の例データ記憶手段に格納された正の例データ群を参照して、前記ペア生成処理手段によって生成されたペアが前記正の例データ群に存在するか否かを判定する存在判定処理手段と、

前記入力データから生成されたペアから前記正の例データ記憶手段に格納された正の例データ群に存在しないペアを抽出し、前記抽出したペアについて、当該ペアの各項が前記正の例データ群に出現する出現確率を算出し、当該各項の出現確率の積にもとづいて前記ペアが前記正の例データ群に出現する出現確率を算出する出現確率算出処理手段と、

前記出現確率算出処理手段によって算出された前記ペアの出現確率をもとに、前記ペアが負の例である傾向を示す負の例度合いを求める負の例度合い算出処理手段と、

前記負の例度合い算出処理手段によって求められた前記負の例度合いが所定より大きいペアを外の関係の連体節として検出する外の関係検出処理手段とを備える

ことを特徴とする負の例予測処理を用いた外の関係検出処理装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、負の例予測処理方法、処理プログラムおよび処理装置、負の例予測処理を用いた日本語表記誤り検出処理プログラムおよび処理装置、ならびに、負の例予測処理を用いた外の関係検出処理プログラムおよび処理装置に関する。

【0002】

本発明は、日本語文の表記誤りや日本語構文解析における格関係の判断等に应用することができる。特に、本発明は、実際の日本語文の表記誤りの検出に役に立ち、日本語ワードプロセッサシステムやOCR読み取りシステムへ適用することができる。

【0003】

【従来の技術】

正の例から負の例を予測する処理について、例えば日本語文の表記誤り検出の問題を考える。この場合に、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、日本語の表記誤り（負の例）を予測し抽出することになる。この意味で、正の例から負の例を予測する処理の実現は、実際の日本語文の表記誤りの検出など、正の例があるが負の例の取得が困難な問題の解決に役に立つ重要な課題である。

【0004】

正の例からの負の例の予測方法について、単純な方法だと、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、精度の高い処理に適用することができない。

【0005】

また、以下の参考文献1でも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。つまり、正の例と負の例の両方を教師信号として用いる機械学習手法であれば高精度な処理を期待することができるが、正の例のみの機械学習法では処理の精度に問題があると考えられる。

[参考文献1：横森貫 他，形式言語の学習 - 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]

このように、従来は、正の例から負の例を予測する処理について、実用が可能な程度に精度が高い処理方法は実現されていなかった。

【0006】

【発明が解決しようとする課題】

本発明は、従来実現されていなかった実用可能な精度を備えた正の例から負の例を予測

10

20

30

40

50

する処理方法、その処理方法をコンピュータに実行させるためのプログラム、および処理装置を提供することを目的とする。

【0007】

さらに、この負の例予測処理方法を用いた日本語表記誤り検出処理をコンピュータに実行させるためのプログラムおよび、その処理装置を提供することを目的とする。

【0008】

さらに、この負の例予測処理方法を用いた格関係において外の関係の文抽出処理をコンピュータに実行させるためのプログラムおよび、その処理装置を提供することを目的とする。

【0009】

【課題を解決するための手段】

本発明は、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する。

【0010】

本発明は、ある問題について正または負であることが未知のデータについて負であるか否かを予測する処理を行うために、前記問題について正の例である正の例データ群を記憶する正の例データ記憶手段と、データ入力処理手段、存在判定処理手段、出現確率算出処理手段、および負の例度合い算出処理手段とを備えたコンピュータが行う処理方法であって、1) 前記データ入力処理手段が、 予測処理の対象となる2つの項で構成されるデータを取得するデータ入力処理過程と、2) 前記存在判定処理手段が、 前記正の例データ記憶手段に格納された正の例データ群を参照して、前記データ入力処理過程において取得された入力データが前記正の例データ群に存在するか否かを判定する存在判定処理過程と、3) 前記出現確率算出処理手段が、 前記入力データが前記正の例データ記憶手段に格納された正の例データ群に存在しない場合に、前記入力データの各項が前記正の例データ記憶手段に格納された正の例データ群に出現する出現確率を算出し、当該入力データの各項の出現確率の積にもとづいて前記入力データが前記正の例データ群に出現する出現確率を算出する出現確率算出処理過程と、4) 前記負の例度合い算出処理手段が、 前記出現確率算出処理過程において算出された入力データの出現確率をもとに、前記入力データが前記問題について負の例である傾向を示す負の例度合いを求める負の例度合い算出処理過程とを備える。

【0011】

本発明は、前記の負の例予測処理方法で実行され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る。また、本発明は、前記の負の例予測処理を実行する処理手段を備える処理装置である。

【0012】

さらに、本発明は、上記の実際の日本語文の表記誤りの検出や、文の格関係について外の関係の文の抽出等に適用することができるが、正の例は存在するが負の例の獲得が困難な種々の問題全般に適用することができる。

【0013】

また、本発明は、日本語表記の正誤検出処理において正または負であることが未知のデータについて日本語表記の誤りであることを示す負の例であるか否かを予測する負の例予測処理を用いて、入力されたデータの日本語表記誤りを検出する処理を、正しい日本語表記の例である正の例データ群を記憶する正の例データ記憶手段を備えたコンピュータに実行させるための負の例予測処理を用いた日本語表記誤り検出処理プログラムであって、1) 前記正の例データ記憶手段にアクセスする正の例データアクセス処理と、 2) 検出処理の対象となるデータを取得するデータ入力処理と、 3) 前記データ入力処理によって取得された入力データの文字の各すき間について、 前記すき間に接続する前接文字列および後

10

20

30

40

50

接文字列を項とするペアを生成するペア生成処理と、4)前記正の例データ記憶手段に格納された正の例データ群を参照して、前記ペア生成処理によって生成されたペアが前記正の例データ群に存在するか否かを判定する存在判定処理と、5)前記入力データから生成されたペアから前記正の例データ記憶手段に格納された正の例データ群に存在しないペアを抽出し、前記抽出されたペアについて、当該ペアの各項が前記正の例データ群に出現する出現確率を算出し、当該ペアの各項の出現確率の積にもとづいて前記ペアが前記正の例データ群に出現する出現確率を算出する出現確率算出処理と、6)前記出現確率算出処理によって算出された前記ペアの出現確率をもとに、前記ペアが負の例である傾向を示す負の例度合いを求める負の例度合い算出処理と、7)前記負の例度合い算出処理によって求められた前記負の例度合いが所定より大きいペアを表記の誤り箇所として検出する表記誤り検出処理とを、コンピュータに実行させるためのものである。

10

【0014】

また、本発明は、前記の負の例予測処理を用いた日本語表記誤り検出処理プログラムの処理を実行する処理手段を備える処理装置である。

【0015】

さらに、本発明は、連体節の格関係が未知のデータについて外の関係を示す負の例であるか否かを予測する負の例予測処理を用いて、入力されたデータの外の関係となる連体節を検出する処理を、格関係にある動詞と名詞との組を持つ正の例データ群を記憶する正の例データ記憶手段を備えたコンピュータに実行させるための負の例予測処理を用いた外の関係検出処理プログラムであって、1)前記正の例データ記憶手段にアクセスする正の例データアクセス処理と、2)検出処理の対象となるデータを取得するデータ入力処理と、3)前記データ入力処理によって取得された入力データの連体修飾関係を構成する連体節の動詞とのかかり先の名詞とを抽出し、前記動詞および前記名詞を項とするペアを生成するペア生成処理と、4)前記正の例データ記憶手段に格納された正の例データ群を参照して、前記ペア生成処理によって生成されたペアが前記正の例データ群に存在するか否かを判定する存在判定処理と、5)前記入力データから生成されたペアから前記正の例データ記憶手段に格納された正の例データ群に存在しないペアを抽出し、前記抽出したペアについて、当該ペアの各項が前記正の例データ群に出現する出現確率を算出し、当該各項の出現確率の積にもとづいて前記ペアが前記正の例データ群に出現する出現確率を算出する出現確率算出処理と、6)前記出現確率算出処理によって算出された前記ペアの出現確率をもとに、前記ペアが負の例である傾向を示す負の例度合いを求める負の例度合い算出処理と、7)前記負の例度合い算出処理において求められた前記負の例度合いが所定より大きいペアを外の関係の連体節として検出する外の関係検出処理とを、コンピュータに実行させるためのものである。

20

30

【0016】

また、本発明は、前記の負の例予測処理を用いた外の関係検出処理プログラムの処理を実行する処理手段を備える処理装置である。

【0017】

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。

40

【0018】

【発明の実施の形態】

図1に、本発明にかかる負の例予測処理装置の構成例を示す。負の例予測処理装置10は、存在判定部11と、出現確率推定部12と、負の例度合い算出部13と、正の例データベース14を持つ。

【0019】

存在判定部11は、入力された未知の事例xが正の例データベース14に存在するかどうかを判定する手段である。

50

【 0 0 2 0 】

出現確率推定部 1 2 は、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。

【 0 0 2 1 】

負の例度合い算出部 1 3 は、一般的な出現確率（頻度） $p(x)$ 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。

【 0 0 2 2 】

正の例データベース 1 4 は、正の例データ D を記憶する記憶手段である。

【 0 0 2 3 】

図 2 に、負の例予測処理の処理フローチャートを示す。

10

【 0 0 2 4 】

まず、存在判定部 1 1 は、正の例か負の例か判定すべき未知の事例 x を入力する（ステップ $S 1$ ）。入力する事例 x は、 (a, b) の二項関係で与えられると仮定する。

【 0 0 2 5 】

存在判定部 1 1 は、入力された未知の事例 x が正の例データ D に含まれるかどうかを調べ（ステップ $S 2$ ）、未知の事例 x が正の例データ D に含まれないときは（ステップ $S 3$ ）、ステップ $S 4$ の処理を行なう。

【 0 0 2 6 】

ステップ $S 4$ では、未知の事例 x の一般的な出現確率 $p(x)$ を推定する。例えば、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 D 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。

20

【 0 0 2 7 】

なお、事例 x の一般的な出現確率 $p(x)$ は、何らかの方法で算出できればよく、上記の方法に限られるものではない。

【 0 0 2 8 】

次に、負の例度合い算出部 1 3 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 D に出現する確率 $Q(x)$ を推定する（ステップ $S 5$ ）。

30

【 0 0 2 9 】

このとき、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 D に出現しない確率は $(1 - p(x))^n$ となり、事例 x が正の例データ D に出現する確率 $Q(x) = 1 - (1 - p(x))^n$ となる。

【 0 0 3 0 】

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 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 D （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。

40

【 0 0 3 1 】

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。

【 0 0 3 2 】

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。

【 0 0 3 3 】

50

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。

【0034】

なお、ステップ $S4$ の処理で、事例 x が正の例データ D のデータベース 14 に含まれるときは、負の例度合い算出部 13 は、事例 x を正の例であると判定し、負の例度合い $Q(x) = 0$ とする(ステップ $S6$)。

【0035】

以上の説明のように、本発明は、正の例データ D の頻度情報を用いて負の例を予測することができ、また、負の例の度合いを数値化して出力することができる。 10

【0036】

次に、本発明の有効性および汎用性を示すため、本発明を日本語表記誤り検出の問題と外の関係の文の抽出の問題とに適用した場合の処理を説明する。

【0037】

[第1の実施の形態：日本語表記誤りの検出処理]

第1の実施の形態として、本発明を日本語表記誤り検出の問題に適用した場合の処理を説明する。

【0038】

単語の表記誤りに限って言えば、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ったとしても扱うのが困難である。 20

【0039】

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。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。

【0040】

まず、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2～参考文献4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と判定する。 30

[参考文献2：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3：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4：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] 40

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5～参考文献7に記載されている。

[参考文献5：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6：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7：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)]

上記にあげた従来手法のうち、参考文献6の $ngram$ 確率を利用する手法は、主に OC 50

R 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。

【0041】

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 7 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。

【0042】

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 T_p 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 T_s 以上となった文字を誤りと判定する（ $T_p = 0$ 、 $T_s = -2$ とする）。ここで、 $T_p = 0$ としているために確率は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。 $T_p > 0$ とした場合は、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 $T_p > 0$ は適切ではなく、 $T_p = 0$ の設定は良いとする。

10

【0043】

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、頭から「負の事」「の事零」といった 3 文字連続を切り出し、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 3 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。

20

【0044】

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。また、コーパスでの確率や頻度を用いないものとなっている点が、コーパスでの確率頻度を用いる本発明と異なる。

30

【0045】

以下に、本発明を表記誤り検出方法に適用した場合の処理を説明する。

【0046】

図 4 に、本発明を適用した表記誤り検出装置 20 の構成例を示す。表記誤り検出装置 20 は、ペア生成部 21 と、ペア管理部 22 と、正負判定部 23 と、正の例データベース 24 とを持つ。

【0047】

ペア生成部 21 は、入力された文章のチェック対象である文字の各すき間から、すき間に接続する前接文字列および後接文字列の二項関係のペア $x(a, b)$ を生成する手段である。

40

【0048】

ペア管理部 22 は、ペア生成部 21 から、ペア $x(a, b)$ で正のデータ D に存在しないと判定されたものを受け取り、正負判定部 23 で算出されるペア x の負の例度合い $Q(x)$ を管理する手段である。

【0049】

正負判定部 23 は、図 1 に示す負の例予測処理装置 10 と同様の処理を行って、ペア管理部 22 から渡されたペア x の負の例度合いを算出する手段である。

【0050】

正負判定部 23 は、図 1 に示す存在判定部 11 と同様の処理によりペア生成部 21 で生成されたペア x が正の例データ D に存在するか否かを判定するペア存在判定部 231 と、同

50

じく出現確率推定部 1 2 と同様の処理によりペア x の一般的な出現確率 $p(x)$ を算出する出現確率推定部 2 3 2 と、同じく負の例度合い算出部 1 3 と同様の処理により、ペア x の負の例度合い $Q(x)$ を算出する負の例度合い算出部 2 3 3 とを備える。

【 0 0 5 1 】

正の例データベース 2 4 は、正の例データ D を記憶するデータベースであり、ここでは正しい日本語文の集合であるコーパスを用いる。

【 0 0 5 2 】

本形態では、処理対象となるペア x の二項関係 (a, b) を、各すき間に接続する任意の連続する 1 ~ 5 g r a m の二つの文字列の関係とする。基本的な考え方は、この二つの文字列 a 、 b の接続チェックを正の例データベース (コーパス) 2 4 で行なうことにより表記の誤りを検出する。二つの文字列 a 、 b が接続できる場合を「正の例」とし、接続できない場合を「負の例」とする。

10

【 0 0 5 3 】

図 5 に、本形態における、表記誤り検出処理の処理フローチャートを示す。

【 0 0 5 4 】

表記誤り検出装置 2 0 のペア生成部 2 1 は、処理対象である文章を入力して (ステップ S 1 1)、すべてのすき間について処理が終了するまで (ステップ S 1 2)、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以下のステップ S 1 4 ~ 1 9 の処理を行なう (ステップ S 1 3)。

【 0 0 5 5 】

ペア生成部 2 1 は、対象としている文字のすき間に前接する 1 ~ 5 g r a m の文字列 a と、後接する 1 ~ 5 g r a m の文字列 b を取り出し、この任意のペア $x = (a, b)$ を作る。ここで、2 5 個のペアが生成されることになる (ステップ S 1 4)。

20

【 0 0 5 6 】

そして、存在判定部 2 3 1 により、ペア x の 2 5 個の接続 $a b$ が正の例データベース 2 4 にあるかどうかを調べ、その判定結果をペア生成部 2 1 に返す (ステップ S 1 5)。

【 0 0 5 7 】

存在判定部 2 3 1 から判定結果を受け取ったペア生成部 2 1 は、ペア x の接続 $a b$ がコーパス 2 4 にあると判定されたペア x を除いた残りのペア x をペア管理部 2 2 に渡す (ステップ S 1 6)。

30

【 0 0 5 8 】

なお、ペア生成部 2 1 は、すべてのペア x がコーパス 2 4 にあるという判定を存在判定部 2 3 1 から受け取った場合は、そのすき間は接続するものと判断し、接続は妥当なもの (正の例) と判定し、負の例度合い $Q(x) = 0$ とし (ステップ S 1 7)、そして処理を次のすき間に移す (ステップ S 1 3)。

【 0 0 5 9 】

ペア管理部 2 2 では、ペア生成部 2 1 から受け取ったペア x について、正負判定部 2 3 に渡し、正負判定部 2 3 の出現確率推定部 2 3 2 は、各ペア x ごとに上記の説明のようにして負の例度合い $Q(x)$ を求める (ステップ S 1 8)。

【 0 0 6 0 】

そして、ペア管理部 2 2 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判定する。そして、処理を次のすき間に移す (ステップ S 1 9)。

40

【 0 0 6 1 】

上記の処理では、二項関係を各箇所 (すき間) で 2 5 種類を作成し、それぞれで負の例度合い $Q(x)$ を求め、 $Q(x)$ のもっとも大きいときの値 $Q(x_{max})$ を最終的な判定に用いているものになっている。すなわち、接続チェックのパターンとして 2 5 種類を用意し、この中でもっとも負の例の度合いの大きいパターンを最終的な評価に利用するということにする。

【 0 0 6 2 】

50

一般的に妥当性のチェックという場合には、各種のチェック機構を用いてチェックを行い、そのうち一つにでも該当するとして検出されるときは妥当でないと判断するのが適当である。本発明でも、かかる妥当性チェックの場合と類似して、多くのチェックパターンを用意して、その中でチェックにかかったところのうち最も大きな値となるチェックの評価を最終評価に用いている。

【0063】

ところで、正の例データDとしては、実際に誤り検出をかけるデータ自身も用いることができる。ここで、自分自身を用いるために、当然自分自身のデータによりチェックの対象となる表現は必ず1回以上検出されることになる。このため、出現頻度は1を引いて用いるようにする。これは、leave one out法と等価である。なお、この場合に

10

【0064】

本発明の有効性を確かめるために行った具体例を説明する。

【0065】

まず、参考文献4に記載されている従来手法(以下、従来手法Bという)で示された誤り例を検出できるかどうかの実験を行なった。図6に、参考文献4で示された9つの誤りを含む例文を示す。例文の下線部分は誤り部分である。正の例データDとして、M新聞の91年から98年の文章データを用いた。

【0066】

図7に、本発明による表記誤り検出処理の結果のうち負の度合いの高い上位10個の事例を示す。図7において、上位では負の例度合いが極めて高く、ほとんど上限の1に近いことがわかる。また、例文8の「意味ネットワーク」以外は、抽出したすべてで表記誤りの検出に成功していることがわかる。また、図6に示す例文については、例文1「自然な(つながりをもつようにする)必要がある。(括弧内は図6に下線部で示す誤り部分を示す)」という一つの事例を除いたすべての事例を上位25個以内に検出できていた。

20

【0067】

なお、コーパスのあらゆるひらがな連続を辞書に登録し、コーパスに無いひらがな連続を誤りとする従来手法Bでは、例文8および例文9が検出できなかった。しかし、本発明によれば例文8および例文9の誤りについても上位で検出することができた。

30

【0068】

また比較のために、従来手法Aについても同様の条件で実施してみた。従来手法Aでは、11箇所を誤り候補として検出した。正しく検出できたものは3例のみであり再現率に問題があると考えられる。

【0069】

次に、作為的に誤り箇所を生成したデータを用いた擬似的な別の具体例について説明する。

【0070】

本例では、京大コーパスにあるM新聞の95年の1月17日までの16日間の約2万文(892,655文字)で行なった。なお、京大コーパスについては、以下の参考文献8に

40

[参考文献8:黒橋禎夫 他,京都大学テキスト・コーパス・プロジェクト,言語処理学会第3回年次大会,(1997),pp.115-118]

本例では、1文字削除、1文字置換、1文字挿入の三種類の作為的な誤りの例についての処理をそれぞれ独立に行なった。

【0071】

また、3種の例のそれぞれにおいて各日に100個の誤りをランダムな箇所に生成し、それぞれ合計1,600個の誤りを作成した。このとき、誤り箇所の前後10文字以内

50

ダムに決定した。

【0072】

作成した誤りが1,600文字で元の文字数が892,655文字であるから、誤り文字の出現率は0.18%で、558文字に1つの割合で誤りが生じていることになる。また、正の例データDとしたものは、M新聞の91年から94年の記事データである。また、処理例は1日分のデータを一つの記事(データ)として入力した。すなわち、上記で説明した自分自身のデータも含めて行なうという方法の自分自身のデータは、この1日分となる。

【0073】

さらに、本発明による処理の他、比較のために従来手法Aによる処理も行なった。図8～図10に、これらの処理結果を示す。図8に1文字削除データでの誤り検出の精度を、図9に1文字置換データでの誤り検出の精度を、図10に1文字挿入データでの誤り検出の精度を示す。ここでは再現率と適合率を評価に用いた。再現率は正解の数を誤りの総数1,600で割ったものを意味し、適合率は正解の数を検出数で割ったものを意味する。図8～図10の「上位X個」は負の度合いQ(x)でソートしたデータの上位X個までについての検出の精度を意味する。

10

【0074】

また、正解の判定は表記誤りをしている1文字を厳密に指摘せず一文字前後にずれて指摘していても正しく検出したと判定する。また、すでに正解不正解の判定をした事例の1文字前後の事例は、その事例の指摘が正解でない場合は以降の判定から除いている。

20

【0075】

図8～図10に示すの検出の精度から以下のことがわかる。

【0076】

当然のことではあるが、上位X個のXが増えるにつれて、すなわち検出数が増えるにつれて再現率が上昇する。上位1,600個のところを見ると、再現率と適合率が一致する。これは誤りの総数と検出数が一致するためである。この時点で調べると、おおよそ、1文字削除データで精度が1/3で(図8参照)、1文字置換/挿入データで精度が1/2であることがわかる(図9および図10参照)。これは、上記したように、本例の擬似的データでは558文字に1つの割合で誤りが生じている状態であるので、おおよそ400字詰原稿用紙1枚半に一つ誤りがあるというときには、1文字削除が約1/3の確率で約1/3を検出でき、1文字置換あるいは挿入がおおよそ半分の確率で半分を検出できることを意味する。なお、一般に誤りの出現率が減ると、誤りでないのに誤りと指摘する誤りが生じて精度は低下する。誤りの出現は正しいものの出現に比べると大幅に小さいので、一般的には単純に誤りの出現率が半分になると、誤った検出になる原因部分が倍になり精度は半分になると考えるとよい。

30

【0077】

次に本発明と従来手法Aと比較する。図8～図10に示すように、従来手法Aでは、誤りの程度を数値化することができない。このために、検出の際にソートする基準となる尺度(値)がなく、検出結果に従って上位だけを抽出して調べるなどといったことができない。

40

【0078】

これに対し、本発明は負の例の度合いを数値として算出することで、検出した誤りの程度を数値化して利用することができる。このため、本発明では、負の例の度合いに基づいて結果をソート処理し、上位の精度よく検出されたところだけを抽出するなどというような後処理を可能とする。

【0079】

そして、後処理として、検出した負の例度合いの大きい箇所、すなわち表記誤りの程度が大きい箇所を表示装置など表示させて、簡単に修正できそうな明らかな誤りを手早く修正することができる。また、負の例度合いをもとに、予め定めた区分けにもとづいた色分け表示、輝度分け表示、ブリンク表示などにより、表記誤りの箇所の表示を他の部分と異な

50

る状態を表示したり、表記誤りの程度自体をグラフ等に表示したりすることができる。

【0080】

また、従来手法Aでは、再現率が固定であり、1文字削除で25%、他のもので60%であり、多くの誤りを必ず見過ごすものとなっているという問題がある。また、基本的な精度についても、検出数が近似した上位5,000くらいで比較すると、本発明の方が高精度の結果を得られている。すなわち、本発明による表記誤り検出で、実用可能な程度の精度を得ていることが分かる。

【0081】

なお、本形態では日本語を対象として処理を説明したが、本発明は、英語などの他の言語における文法エラーチェックなどにも適用することができる。

10

【0082】

〔第2の実施の形態：外の関係の文の抽出〕

第2の実施の形態として、本発明を外の関係の文の抽出の問題に適用した場合の処理を説明する。

【0083】

外の関係の文とは、連体修飾節の動詞と被修飾要素の名詞とが格関係にない文のことをいい、埋め込み文の節の動詞とその係り先の名詞の間に格関係が成立しないものをいう。

「負の事例を抽出することは難しい。」

上記のような文の場合に、「負の事例を抽出すること」という関係節では、「抽出する」という動詞とその係り先の「こと」という名詞の間で、「ことが抽出する」や「ことを抽出する」などのような格関係が成立しない。すなわち、「抽出する」と「こと」の間にガ格やヲ格などの格関係がないために、外の関係の文とされる。これとは逆に格関係が成立する文は、内の関係の文と呼ばれる。

20

【0084】

外の関係の文は上記のような形式的なもの他に、

「さんまを焼くけむり。」

などといった複雑な構造をしたものもある。

【0085】

ここで、格関係にある連体修飾節を正の例とするとき、外の関係の文は負の例となる。格関係にある用言（動詞など）と名詞はコーパス中に多く存在する。このため、本発明により、この情報を正の例として負の例の外の関係の文を予測すると、正の例として格各関係にある動詞と名詞から、負の例としての外の関係の文が自動的に抽出できる。

30

【0086】

外の関係の文を抽出する従来手法として、以下の参考文献9～参考文献11に記載されている手法がある。

【0087】

参考文献9の従来手法は、連体修飾関係と格関係で、それを構成する動詞の異なり数の分布に大きな違いがあることに着目し、その分布の違いをKL-距離を用いて評価することで外の関係の文を特定するものである。また、参考文献10には、連体節に関して外の関係になりやすい名詞をあらかじめ抽出するなどしてその情報を利用した人手ルールに基づく方法を用いた研究から、格フレーム情報を含む広範な情報を属性とした教師あり機械学習を用いて外関係を特定する手法が記載されている。また、参考文献11の手法は、埋め込み文の日英翻訳のために格フレームの情報を用いて外関係か内関係を判定するものである。

40

[参考文献9：阿部川武 他，統計情報を利用した日本語連体修飾節の解析，言語処理学会年次大会，(2001)，pp.269-272]

[参考文献10：Timothy Baldwin, Making lexical sense of japanese-english machine translation: A disambiguation extravaganza, Technical report, (Tokyo Institute of Technology, 2001), Technical Report, ISSN 0918-2802]

[参考文献11：表克次，埋め込み文の日英翻訳方式、鳥取大学卒業論文，(2001)]

50

図 1 1 に、本形態において本発明を適用する外の関係の文抽出装置 3 0 の構成例を示す。外の関係の文抽出装置 3 0 は、図 4 に示す表記誤り検出装置 2 0 と同様の構成であり、ペア生成部 3 1、ペア管理部 3 2、および正負判定部 3 3 は、表記誤り検出装置 2 0 のペア生成部 2 1、ペア管理部 2 2、および正負判定部 2 3 と同等の処理を行う。正の例データベース 3 4 は、正しい日本語文の集合であるコーパスから、構文解析システム (k n p) などを用いて取り出した格関係にあるとされる名詞と動詞の対のデータを正の例データ D として記憶したデータベースである。k n p については、以下の参考文献 1 2 に記載されている。

[参考文献 1 2 : 黒橋禎夫, 日本語構文解析システム K N P 使用説明書 ver.2.0b6]

外の関係の文抽出装置 3 0 では、処理対象 x の二項関係 (a , b) は名詞と動詞の対とする。外の関係の文抽出装置 3 0 は、処理対象 x が高頻度に出現する名詞と動詞の対であるにも関わらず、正の例データ D に存在しなければ、それらは外の関係であろうと判定する。

【 0 0 8 8 】

外の関係の文抽出処理の処理の流れは、図 5 に示す表記誤り検出処理の処理フローチャートに示す処理の流れとほぼ同様である。

【 0 0 8 9 】

まず、コーパスから k n p などを用いて大量の格関係にあるとされる名詞と動詞との組 y を取り出し、組 y は正の例データ D として正の例データベース 3 4 へ記憶されているとする。

【 0 0 9 0 】

そして、外の関係の文抽出処理装置 3 0 のペア生成部 3 1 は、まず、コーパスなどから k n p などを用いて大量の連体節の動詞とのかかり先の名詞との組 $x = (a , b)$ を取り出す。これらのデータが外の関係かどうか判定されるものとなる。

【 0 0 9 1 】

正負判定部 3 3 の存在判定部 3 3 1 は、ペア生成部 3 1 で生成した組 $x = (a , b)$ が組 y の集合すなわち正の例データ D に含まれるか否かを判定する。組 x が正の例データ D に含まれる場合には、存在判定部 3 3 1 は、組 x を正の例と判定し、ペア生成部 3 1 は、この組 x を外の関係 (負の例) でなく内の関係 (正の例) であると判断する。

【 0 0 9 2 】

一方、組 x が正の例データ D (組 y の集合) に含まれない場合は、組 x が名詞と動詞の二項関係からなるものと考えて、その組 x をペア管理部 3 2 へ渡し、ペア管理部 3 2 は、その組 x を正負判定部 3 3 へ渡し、組 x の負の例度合い $Q (x)$ を取得して管理する。

【 0 0 9 3 】

正負判定部 3 3 の出現確率推定部 3 3 2 および負の例度合い算出部 3 3 3 は、上記で説明したような処理により、組 x の負の例度合い $Q (x)$ を算出する。ペア管理部 3 2 は、この負の例度合い $Q (x)$ の値が大きいほど、負の例の度合いが大きいと判定し、外の関係である可能性は高いと判定する。

【 0 0 9 4 】

本発明の有効性を確かめるために行った具体例を説明する。

【 0 0 9 5 】

本例では、少量のデータ (1 , 5 3 0 個) のうちの連体節にかかわるデータ (8 7 0 事例) を用いて行なった。使用するデータでは、各事例が外の関係であるか否かの情報が付与されているために自動的に精度を求めることができる。使用したデータのうち、外の関係の事例は 2 6 7 個であった。なお、正の例データベース 3 4 としては M 新聞の 9 5 年を除く 9 1 年から 9 8 年までの 7 年分の記事データを用いた。

【 0 0 9 6 】

図 1 2 に、本例における検出精度を示す。評価は再現率と適合率と正解率で行なった。再現率は正しく外の関係を特定できた数を外の関係の総数 2 6 7 で割ったものを意味し、適合率は正しく外の関係を特定できた数を検出数で割ったものを意味する。正解率は、その

10

20

30

40

50

正解率を求める地点までの事例を外の関係と判断した場合の全事例 870 個での外の関係と内の関係の区別の正解精度である。

【0097】

図 12 中、「上位 X 個」は負の例度合い $Q(x)$ でソートしたデータの上位 X 個までについての検出精度を意味する。本例での検査精度は、上位 10 個まででは精度は 100% であり、正の例だけからでも、それなりに外の関係の文を抽出できることがわかる。

【0098】

また、全般的に精度が低いとはいえ、上位での適合率は高い。外関係の出現率は 30.7% であり、本例では上位 10 個で連続して正解しているが 30.7% の確率のものを 10 回連続生じる確率は 0.0000074 であり、これは偶然生じるようなことではない。上位の適合率の高さからも、本発明による外関係の文の抽出は、十分に実用可能な程度の精度を得ていると考えられる。

10

【0099】

以上、本発明を、日本語表記誤り検出問題と外関係の文の抽出問題に適用する形態を説明した。両方の問題においても、負の例度合いでソートした結果の上位では高い適合率で負の例を検出でき、本発明の有効性を確認した。また、これらの二つの問題で有効性を確認できたことにより、本発明の汎用性も確認することができた。すなわち、本発明が、他の多くの正の例からの負の例を予測する問題の解決手法として有効であり、これらの問題を同様に解くことができると考えられる。

【0100】

以上、本発明をその実施の態様により説明したが、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ことは当然である。

20

【0101】

【発明の効果】

以上説明したように、本発明により、従来実現されていなかった正の例から負の例を予測する処理方法を提供することが可能となった。

【0102】

本発明は、上記の日本語文の表記誤りの検出、文の格関係について外関係の文の抽出等以外にも、多くの正の例からの負の例を予測する他の問題についても適用が可能であり、かかる問題解決において実用可能な程度の精度を備えた解決手法を提供するという効果を奏する。

30

【0103】

また、本発明は、負の例の予測処理の結果である負の例度合いを数値として出力することができるため、処理結果を種々の後処理に利用することが可能な負の例予測処理を提供するという効果を奏する。

【図面の簡単な説明】

【図 1】本発明にかかる負の例予測処理装置の構成例を示す図である。

【図 2】負の例予測処理の処理フローチャート図である。

【図 3】従来手法 A の補足説明のための図である。

【図 4】本発明を適用した表記誤り検出装置の構成例を示す図である。

40

【図 5】表記誤り検出処理の処理フローチャート図である。

【図 6】誤りを含む例文を示す図である。

【図 7】表記誤り検出処理の結果を示す図である。

【図 8】表記誤り検出処理の精度を示す図である。

【図 9】表記誤り検出処理の精度を示す図である。

【図 10】表記誤り検出処理の精度を示す図である。

【図 11】本発明を適用する外関係の文抽出装置の構成例を示す図である。

【図 12】外関係の文抽出処理の精度を示す図である。

【符号の説明】

10 負の例予測処理装置

50

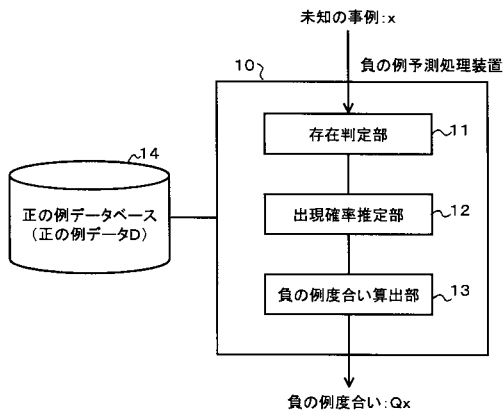
- 1 1 存在判定部
- 1 2 出現確率推定部
- 1 3 負の例度合い算出部
- 1 4 正の例データベース (正の例データ D)
- 2 0 表記誤り検出装置
- 2 1 ペア生成部
- 2 2 ペア管理部
- 2 3 正負判定部
- 2 4 正の例データベース (正の例データ D)
- 2 3 1 存在判定部
- 2 3 2 出現確率推定部
- 2 3 3 負の例度合い算出部
- 3 0 外の関係の文抽出装置
- 3 1 ペア生成部
- 3 2 ペア管理部
- 3 3 正負判定部
- 3 4 正の例データベース (正の例データ D)
- 3 3 1 存在判定部
- 3 3 2 出現確率推定部
- 3 3 3 負の例度合い算出部

10

20

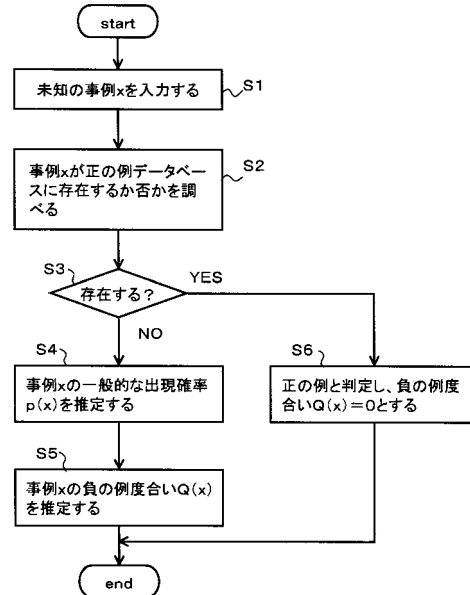
【 図 1 】

負の例予測処理装置の構成例



【 図 2 】

負の例予測処理の処理フローチャート



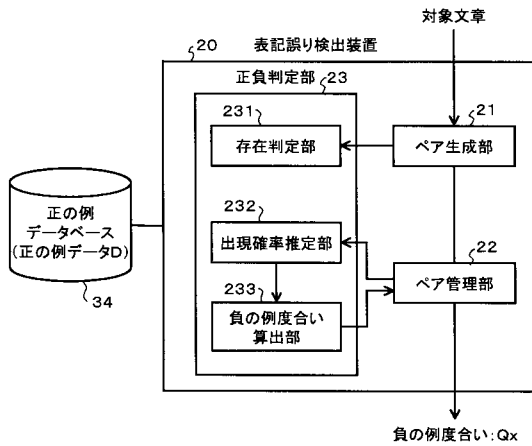
【 図 3 】

例文	負の事零の検出			
trigramによる得点	-1	-1	-1	
		-1	-1	-1
合計得点	-1	-2	-2	-1

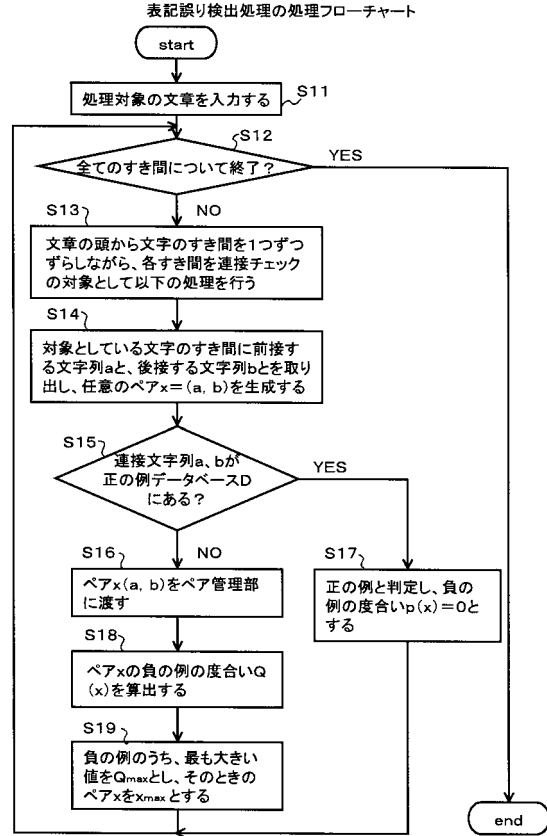
誤り

【 図 4 】

表記誤り検出装置の構成例



【 図 5 】



【 図 6 】

誤りを含む例文

1	自然なつながりをもつようにする必要がある。
2	「～する」「～した」などの表現が用いられる。
3	例えば、図のようなネットワークから、
4	「咲く」を含む文での次ような対応関係を
5	可能な連体節がである場合は、この連体節を
6	説明した方法でを用いることができる
7	文として出力する方が適切であると考え、
8	主題と述語が与えられた場合に意味ネットワーク
9	可能な場合にはすべての並列節として出力した

【 図 7 】

事例	負の例度合い	前方文脈	二項関係	後方文脈
1	0.99999	説明した方	法で	後方支脈
2	0.99999	した」などの裏記が用	いら	用いることができる
3	0.99999	咲く」を含む文での	次	。な対応関係を
4	0.99999	主題と述語が与え	ら	合に意味ネットワーク
5	0.99999	して出力する方が適切	が	考え、であると考え、
6	0.99999	文として出力する方が	適切	のネットワークから
7	0.99999	例えば、	図	ネットワーク
8	0.99999	語が与えられた場合に意	味	力した
9	0.99999	合にはすべての並列節	とい	力した
10	0.99998	合にはすべての並列節	と	

【 図 8 】

		再現率	適合率
本発明の結果			
上位	50 個	2.88%	92.00%
上位	100 個	5.31%	85.00%
上位	200 個	9.69%	77.50%
上位	300 個	12.88%	68.67%
上位	500 個	18.56%	59.40%
上位	800 個	24.06%	48.12%
上位	1,200 個	29.12%	38.83%
上位	1,600 個	32.38%	32.38%
上位	3,000 個	39.75%	21.20%
上位	5,000 個	47.38%	15.16%
上位	10,000 個	57.38%	9.18%
上位	20,000 個	67.81%	5.42%
上位	50,000 個	81.38%	2.60%
上位	100,000 個	89.31%	1.43%
従来手法Aの結果			
総検出数	5,295 個	25.31%	7.65%

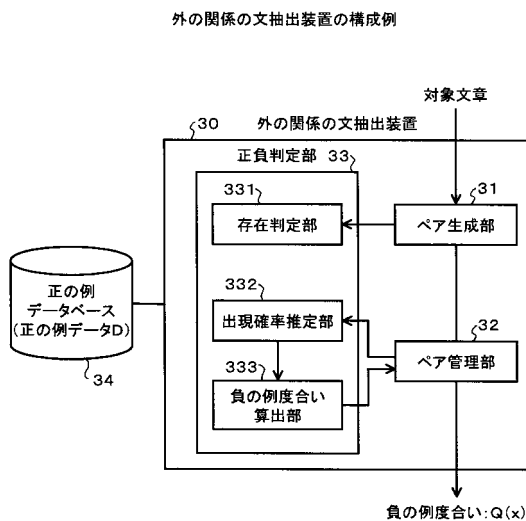
【 図 9 】

		再現率	適合率
本発明の結果			
上位	50 個	3.12%	100.00%
上位	100 個	5.94%	95.00%
上位	200 個	11.56%	92.50%
上位	300 個	16.62%	88.67%
上位	500 個	25.25%	80.80%
上位	800 個	34.25%	68.50%
上位	1,200 個	42.44%	56.58%
上位	1,600 個	48.06%	48.06%
上位	3,000 個	61.38%	32.73%
上位	5,000 個	70.81%	22.66%
上位	10,000 個	81.12%	12.98%
上位	20,000 個	87.62%	7.01%
上位	50,000 個	94.44%	3.02%
上位	100,000 個	97.81%	1.57%
従来手法Aの結果			
総検出数	5,944 個	60.94%	16.40%

【 図 1 0 】

		再現率	適合率
本発明の結果			
上位	50 個	3.12%	100.00%
上位	100 個	6.00%	96.00%
上位	200 個	11.62%	93.00%
上位	300 個	16.69%	89.00%
上位	500 個	24.88%	79.60%
上位	800 個	33.88%	67.75%
上位	1,200 個	41.94%	55.92%
上位	1,600 個	47.19%	47.19%
上位	3,000 個	60.44%	32.23%
上位	5,000 個	69.88%	22.36%
上位	10,000 個	80.62%	12.90%
上位	20,000 個	88.69%	7.09%
上位	50,000 個	95.25%	3.05%
上位	100,000 個	98.12%	1.57%
従来手法Aの結果			
総検出数	5,944 個	62.12%	16.72%

【 図 1 1 】



【 図 1 2 】

		再現率	適合率	正解率
上位	10 個	3.75%	100.00%	70.46%
上位	20 個	7.12%	95.00%	71.38%
上位	30 個	8.99%	80.00%	71.38%
上位	40 個	10.86%	72.50%	71.38%
上位	50 個	13.48%	72.00%	71.84%
上位	100 個	20.97%	56.00%	70.69%
上位	150 個	28.46%	50.67%	69.54%
上位	200 個	33.33%	44.50%	66.78%
上位	300 個	39.70%	35.33%	59.20%
総検出数	393 個	42.32%	28.75%	50.11%

フロントページの続き

(56)参考文献 森中雄, 外3名著, “事例間の類似性に基づく負例の生成によるデータベースからの知識獲得”
 , 人工知能学会誌, 2000年 9月 1日, 第15巻, 第5号, p.862~870.

(58)調査した分野(Int.Cl., DB名)

G06N 5/04

G06F 17/21

G06F 17/27

(54)【発明の名称】負の例予測処理方法、処理プログラムおよび処理装置、負の例予測処理を用いた日本語表記誤り
 検出処理プログラムおよび処理装置、ならびに負の例予測処理を用いた外の関係検出処理プログラ
 ムおよび処理装置