(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2003－196094
(P2003－196094A)
(43)公開日　平成15年７月11日(2003.7.11)

(51)Int.Cl.[7]		識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ	9/44	５５０	Ｇ０６Ｆ 9/44	５５０Ｎ	５Ｂ００９
	17/21	５５０	17/21	５５０Ａ	５Ｂ０６４
	17/27		17/27	Ｘ	５Ｂ０９１
Ｇ０６Ｋ	9/03		Ｇ０６Ｋ 9/03	Ｚ	

審査請求　有　　請求項の数10　ＯＬ　（全 13 頁）

(21)出願番号　特願2001－394112(P2001－394112)

(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　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

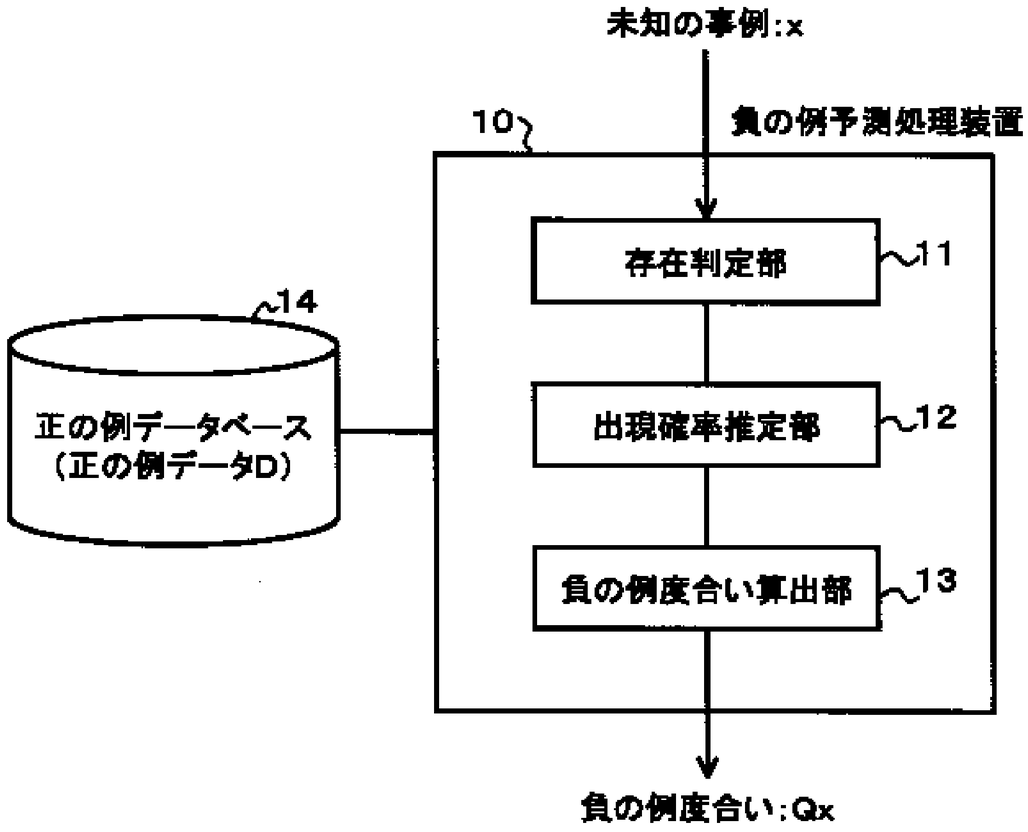
(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１
(72)発明者　村田　真樹
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(74)代理人　100119161
弁理士　重久　啓子　（外１名）

最終頁に続く

(54)【発明の名称】　負の例予測処理方法、負の例予測処理プログラム、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、負の例予測処理を用いた日本語表記誤り検出装置、負の例予測処理を用いた外

(57)【要約】
【課題】　ある問題について正または負の例であることが未知のデータについて負の例である度合いを予測する処理方法を提供する。
【解決手段】　存在判定部１１は、未知の事例ｘを入力し、事例ｘが正の例データベース１４に存在するか否かを判定する。事例ｘが存在しなければ、出現確率推定部１２は、事例ｘの一般的な出現確率ｐ（ｘ）を算出する。負の例度合い算出部１３は、出現確率ｐ（ｘ）から事例ｘが負の例である度合いＱ（ｘ）を算出する。

負の例予測処理装置の構成例



【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が、負であるか否かを予測する処理方法であって、
予め、前記問題について正の例となる正の例データ群を記憶する正の例データ記憶部を備え、
前記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、
前記データが前記正の例データ群に存在しない場合に、前記データの一般的な出現確率を算出する出現確率算出処理過程と、
前記一般出現確率をもとに、前記データが前記正の例データ群に出現する確率を算出し、当該確率を負の例度合いとす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。
【請求項２】　請求項１に記載する負の例予測処理方法において、
さらに、前記正の例データ群に存在しない前記データを、当該データの負の例度合いをもとに降順もしくは昇順に並べ替えて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負の例予測処理方法。
【請求項３】　請求項１に記載する負の例予測処理方法において、
さらに、前記データを、当該データの負の例度合いに対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、異なる表示形態で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負の例予測処理方法。
【請求項４】　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が、負であるか否かを予測する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であって、
前記問題について正の例となる正の例データ群を予め記憶する正の例データ記憶部にアクセスする正の例アクセス処理と、
前記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、
前記データが前記正の例データ群に存在しない場合に、前記データの一般的な出現確率を算出する出現確率算出処理と、
前記一般出現確率をもとに、前記データが前記正の例データ群に出現する確率を算出し、当該確率を負の例度合いとする負の例度合い算出処理とを、
コンピュータに実行させることを特徴とする負の例予測処理プログラム。
【請求項５】　請求項４に記載する負の例予測処理プログラムにおいて、
さらに、前記正の例データ群に存在しない前記データを、当該データの負の例度合いをもとに降順もしくは昇順に並べ替えて出力する負の例出力処理を、
コンピュータに実行させることを特徴とする負の例予測処理プログラム。
【請求項６】　請求項４に記載する負の例予測処理プログラムにおいて、
さらに、前記データを、当該データの負の例度合いに対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、異なる表示形態で表示する負の例表示処理を、
コンピュータに実行させることを特徴とする負の例予測処理プログラム。
【請求項７】　負の例を予測する処理方法を用いて日本語表記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、
予め、日本語表記の正しいデータとなる正の例データ群を記憶する正の例データ記憶部にアクセスする正の例データアクセス処理と、
前記入力された表記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、
前記入力表記が前記正の例データ群に存在しない場合に、前記入力表記の一般的な出現確率を算出する出現確率算出処理と、
前記一般的な出現確率をもとに、前記入力表記が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとする負の例度合い算出処理とを、
コンピュータに実行させ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プログラム。
【請求項８】　負の例を予測する処理方法を用いて日本語表記誤りを検出する処理装置であって、
日本語表記の正しいデータとなる正の例データ群を記憶する正の例データ記憶手段と、
前記入力された表記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、
前記入力表記が前記正の例データ群に存在しない場合に、前記入力表記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、
前記一般的な出現確率をもとに、前記入力表記が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとす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を用いた日本語表記誤り検出装置。
【請求項９】　負の例を予測する処理方法を用いて外の関係の文となる連体節を抽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、
予め、内の関係の文である正の例データ群を記憶する正の例データ記憶部にアクセスする正の例データアクセス処理と、
入力された連体節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、
前記連体節が前記正の例データ群に存在しない場合に、前記連体節の一般的な出現確率を算出する出現確率算出

処理と、
前記一般的な出現確率をもとに、前記連体節が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとする負の例度合い算出処理とを、
コンピュータに実行させることを特徴とする負の例予測処理を用いた外の関係の文抽出処理プログラム。
【請求項１０】　負の例を予測する処理方法を用いて外の関係の文となる連体節を抽出する処理装置であって、
予め、内の関係の文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、
入力された連体節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、
前記連体節が前記正の例データ群に存在しない場合に、前記連体節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、
前記一般的な出現確率をもとに、前記連体節が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとす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を用いた外の関係の文抽出装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、正の例から負の例を予測する処理方法、および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、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その処理装置、並びに、負の例予測処理を用いた外の関係の文抽出処理プログラムおよびその処理装置に関する。
【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応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の負の例を予測する処理方法、および、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が、負であるか否かを予測する処理方法であって、予め、前記問題について正の例となる正の例データ群を記憶する正の例データ記憶部を備え、前記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、前記データが前記正の例データ群に存在しない場合に、前記データの一般的な出現確率を算出する出現確率算出処理過程と、前記一般出現確率をもとに、前記データが前記正の例データ群に出現する確率を算出し、当該確率を負の例度合いとする負の例度合い算出処理過程とを備える。
【００１１】また、本発明は、上記の負の例予測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】本発明は、負の例を予測する処理方法を用いて日本語表記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、予め、日本語表記の正しいデータとなる正の例データ群を記憶する正の例データ記憶部にアクセスする正の例データアクセス処理と、前記入力された表記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、前記入力表記が前記正の例データ群に存在しない場合に、前記入力表記の一般的な出現確率を算出する出現確率算出処理と、前記一般的な出現確率をもとに、前記入力表記が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとする負の例度合い算出処理とを、コンピュータに実行させるものである。

【００１４】また、本発明は、負の例を予測する処理方法を用いて日本語表記誤りを検出する処理装置であって、日本語表記の正しいデータとなる正の例データ群を記憶する正の例データ記憶手段と、前記入力された表記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、前記入力表記が前記正の例データ群に存在しない場合に、前記入力表記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、前記一般的な出現確率をもとに、前記入力表記が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとする負の例度合い算出処理手段とを備える。

【００１５】さらに、本発明は、負の例を予測する処理方法を用いて外の関係の文となる連体節を抽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、予め、内の関係の文である正の例データ群を記憶する正の例データ記憶部にアクセスする正の例データアクセス処理と、入力された連体節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、前記連体節が前記正の例データ群に存在しない場合に、前記連体節の一般的な出現確率を算出する出現確率算出処理と、前記一般的な出現確率をもとに、前記連体節が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとする負の例度合い算出処理とを、コンピュータに実行させるものである。

【００１６】また、本発明は、負の例を予測する処理方法を用いて外の関係の文となる連体節を抽出する処理装置であって、予め、内の関係の文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、入力された連体節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、前記連体節が前記正の例データ群に存在しない場合に、前記連体節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、前記一般的な出現確率をもとに、前記連体節が前記正の例データ群に出現する確率を算出して、当該確率を負の度合いとする負の例度合い算出処理手段とを備える。

【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。

【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。

【００２０】出現確率推定部１２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。

【００２１】負の例度合い算出部１３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。

【００２２】正の例データベース１４は、正の例データＤを記憶する記憶手段である。

【００２３】図２に、負の例予測処理の処理フローチャートを示す。

【００２４】まず、存在判定部１１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Ｓ１）。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。

【００２５】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Ｄに含まれるかどうかを調べ（ステップＳ２）、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Ｓ３）、ステップＳ４の処理を行なう。

【００２６】ステップＳ４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）×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。

【００２７】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。

【００２８】次に、負の例度合い算出部１３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Ｓ５）。

【００２９】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回連続して起こるということから、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。

【００３０】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。

【００３１】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。

【００３２】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。

【００３３】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。

【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いえば、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。

【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、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、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と判定する。

［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ＯＣＲ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（Ｔｐ＝０、Ｔｓ＝－２とする）。ここで、Ｔｐ＝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。Ｔｐ＞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Ｔｐ＞０は適切ではなく、Ｔｐ＝０の設定は良いとする。

【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。

【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。

【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、すき間に連接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。

【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、同じく出現確率推定部１２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２３２と、同じく負の例度合い算出部１３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２３３とを備える。

【００５１】正の例データベース２４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。

【００５２】本形態では、処理対象となるペアｘの二項関係（ａ，ｂ）を、各すき間に連接する任意の連続する１～５ｇｒａｍ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の連接チェックを正の例データベース（コーパス）２４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が連接できる場合を「正の例」とし、連接できない場合を「負の例」とする。

【００５３】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。

【００５４】表記誤り検出装置２０のペア生成部２１は、処理対象である文章を入力して（ステップＳ１１）、すべてのすき間について処理が終了するまで（ステップＳ１２）、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以下のステップＳ１４～１９の処理を行なう（ステップＳ１３）。

【００５５】ペア生成部２１は、対象としている文字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ｘ＝（ａ、ｂ）を作る。ここで、２５個のペアが生成されることになる（ステップＳ１４）。

【００５６】そして、存在判定部２３１により、ペアｘの２５個の連接ａｂが正の例データベース２４にあるかどうかを調べ、その判定結果をペア生成部２１に返す（ステップＳ１５）。

【００５７】存在判定部２３１から判定結果を受け取ったペア生成部２１は、ペアｘの連接ａｂがコーパス２４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２２に渡す（ステップＳ１６）。

【００５８】なお、ペア生成部２１は、すべてのペアｘがコーパス２４にあるという判定を存在判定部２３１から受け取った場合は、そのすき間は連接するものと判断し、連接は妥当なもの（正の例）と判定し、負の例度合いＱ（ｘ）＝０とし（ステップＳ１７）、そして処理を

次のすき間に移す（ステップＳ１３）。
【００５９】ペア管理部２２では、ペア生成部２１から受け取ったペアｘについて、正負判定部２３に渡し、正負判定部２３の出現確率推定部２３２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Ｓ１８）。
【００６０】そして、ペア管理部２２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判定する。そして、処理を次のすき間に移す（ステップＳ１９）。
【００６１】上記の処理では、二項関係を各箇所（すき間）で２５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、連接チェックのパターンとして２５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。
【００６２】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、正の例データＤの全てのデータを通じて二回以上まったく同じ誤りが出現するときは、その誤りは検出できないという問題があるので、検出結果の利用には注意が必要である。
【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（つながりがもつようにする）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。
【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。
【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、以下の参照文献８に説明されている。
［参照文献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文字以内に他の誤りが出現しないような条件を設けた。また、置換、挿入時に新たに置かれる文字は、京大コーパスの９１年から９４年のデータでの文字の出現頻度分布に比例する条件でランダムに決定した。
【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。

【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に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一文字前後の事例は、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。

【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。

【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。

【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る状態で表示したり、表記誤りの程度自体をグラフ等で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。

【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。

【００８４】外の関係の文は上記のような形式的なものの他に、

「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。

【００８６】外の関係の文を抽出する従来手法として、以下の参照文献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の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の関係か内の関係かを判定するものである。

［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)］

図１１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３０の構成例を示す。外の関係の文抽出装置３０は、図４に示す表記誤り検出装置２０と同様の構成であり、ペア生成部３１、ペア管理部３２、および正負判定部３３は、表記誤り検出装置２０のペア生成部２１、ペア管理部２２、および正負判定部２３と同等の処理を行う。正の例データベース３４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（ｋｎｐ）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。ｋｎｐについては、以下の参考文献１２に記載されている。

［参考文献１２：黒橋禎夫，日本語構文解析システムＫＮＰ使用説明書　ver.2.0b6 ］

外の関係の文抽出装置３０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３０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。

【００８８】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。

【００８９】まず、コーパスからｋｎｐ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３４へ記憶されているとする。

【００９０】そして、外の関係の文抽出処理装置３０のペア生成部３１は、まず、コーパスなどからｋｎｐなどを用いて大量の連体節の動詞とそ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。

【００９１】正負判定部３３の存在判定部３３１は、ペア生成部３１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３３１は、組ｘを正の例と判定し、ペア生成部３１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。

【００９２】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３２へ渡し、ペア管理部３２は、その組ｘを正負判定部３３へ渡し、組ｘの負の例度合いＱ（ｘ）を取得して管理する。

【００９３】正負判定部３３の出現確率推定部３３２および負の例度合い算出部３３３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３２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。

【００９４】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。

【００９５】本例では、少量のデータ（１，５３０個）のうちの連体節にかかわるデータ（８７０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２６７個であった。なお、正の例データベース３４としてはM新聞の９５年を除く９１年から９８年までの７年分の記事データを用いた。

【００９６】図１２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２６７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。外の関係の出現率は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、本発明による外の関係の文の抽出は、十分に実用可能な程度の精度を得ていると考えられる。

【００９９】以上、本発明を、日本語表記誤り検出問題と外の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。

【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の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。

【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。

【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。

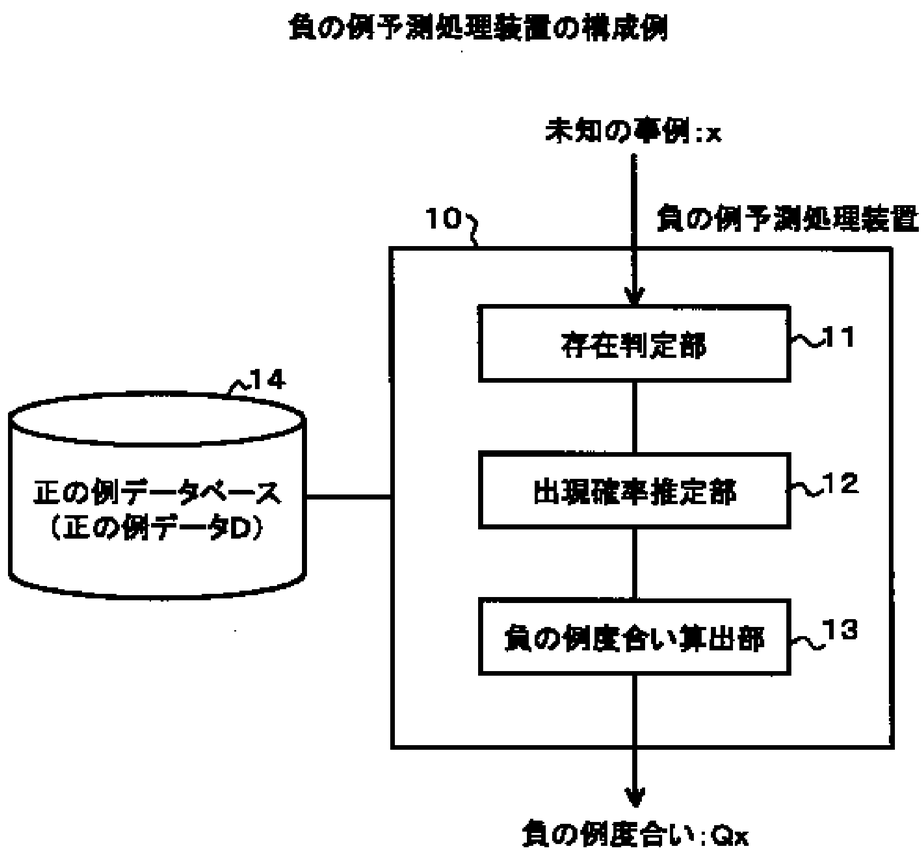
【図１１】本発明を適用する外の関係の文抽出装置の構成例を示す図である。

【図１２】外の関係の文抽出処理の精度を示す図である。

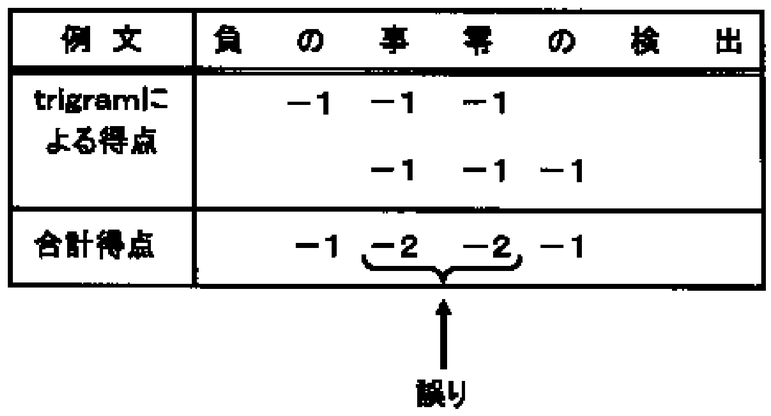
【符号の説明】

１０　負の例予測処理装置
１１　存在判定部
１２　出現確率推定部
１３　負の例度合い算出部
１４　正の例データベース（正の例データＤ）
２０　表記誤り検出装置
２１　ペア生成部
２２　ペア管理部
２３　正負判定部
２４　正の例データベース（正の例データＤ）
２３１　存在判定部
２３２　出現確率推定部
２３３　負の例度合い算出部
３０　外の関係の文抽出装置
３１　ペア生成部
３２　ペア管理部
３３　正負判定部
３４　正の例データベース（正の例データＤ）
３３１　存在判定部
３３２　出現確率推定部
３３３　負の例度合い算出部

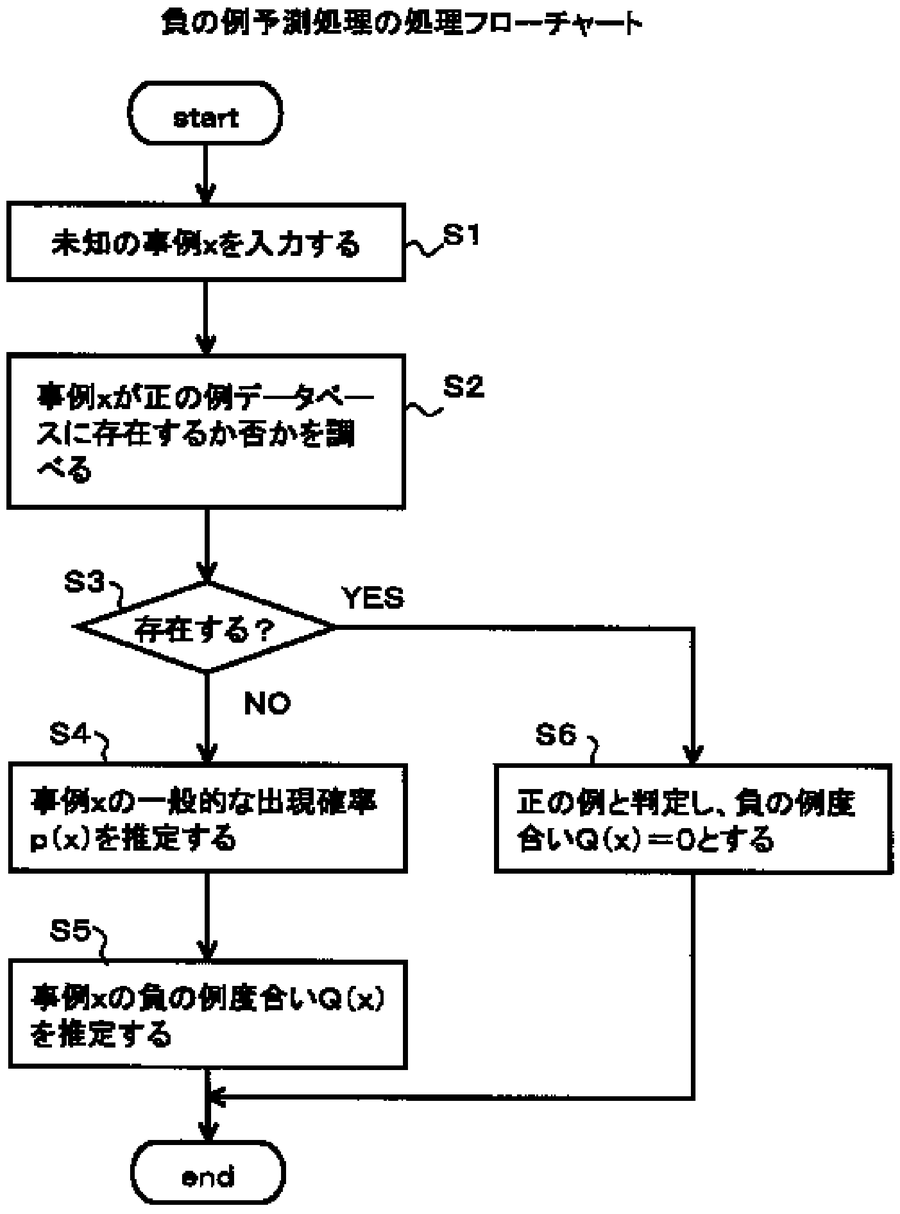
【図１】



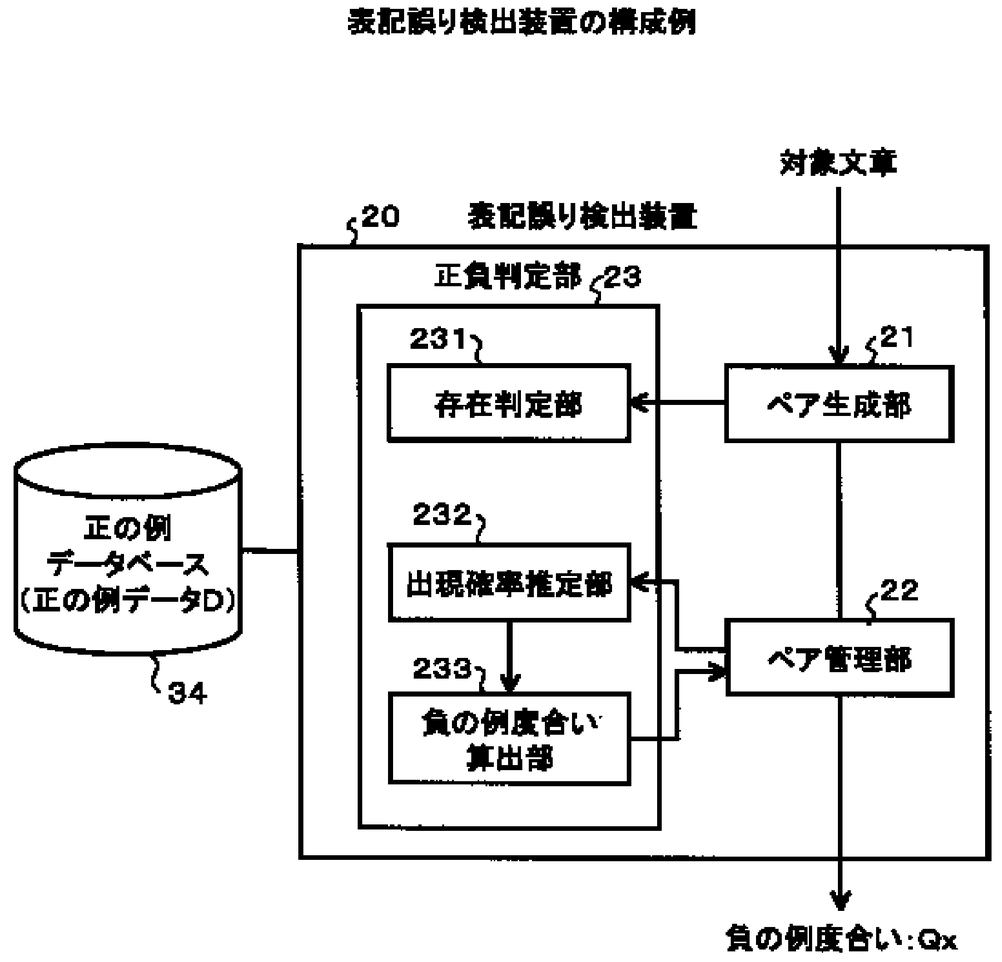
【図３】



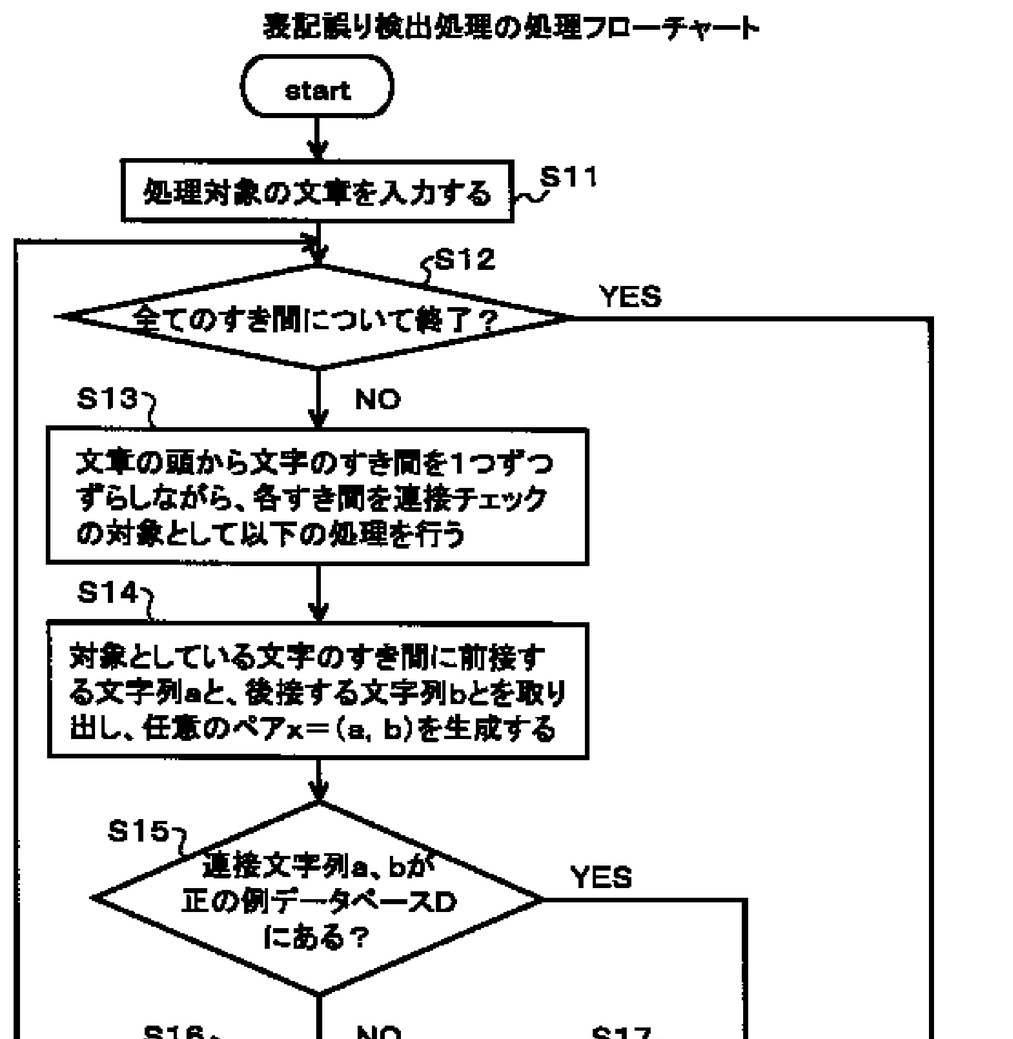
【図２】



【図４】



【図５】



誤りを含む例文

1	自然なつながりが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る。
3	例えば、図ようなのネットワークから、
4	「咲く」を含む文でのは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がであると考え、
8	主題と述語が与えら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いて出力した

【図１２】

		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%

【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	を	用いることができる
2	0.99999	した」などの表記が用	いら	る	。
3	0.99999	咲く」を含む文でのは	次	よう	な対応関係を
4	0.99999	主題と述語が与え	ら	た場	合に意味ネットワーク
5	0.99999	して出力する方が適切	が	であると	考え、
6	0.99999	文として出力する方が	適切	が	であると考え、
7	0.99999	例えば、	図	よう	なのネットワークから
8	0.99999	語が与えられた場合に意	味	ネ	ットワーク
9	0.99999	合にはすべての並列節	とい	て出	力した
10	0.99998	合にはすべての並列節	と	いて出	力した

【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】

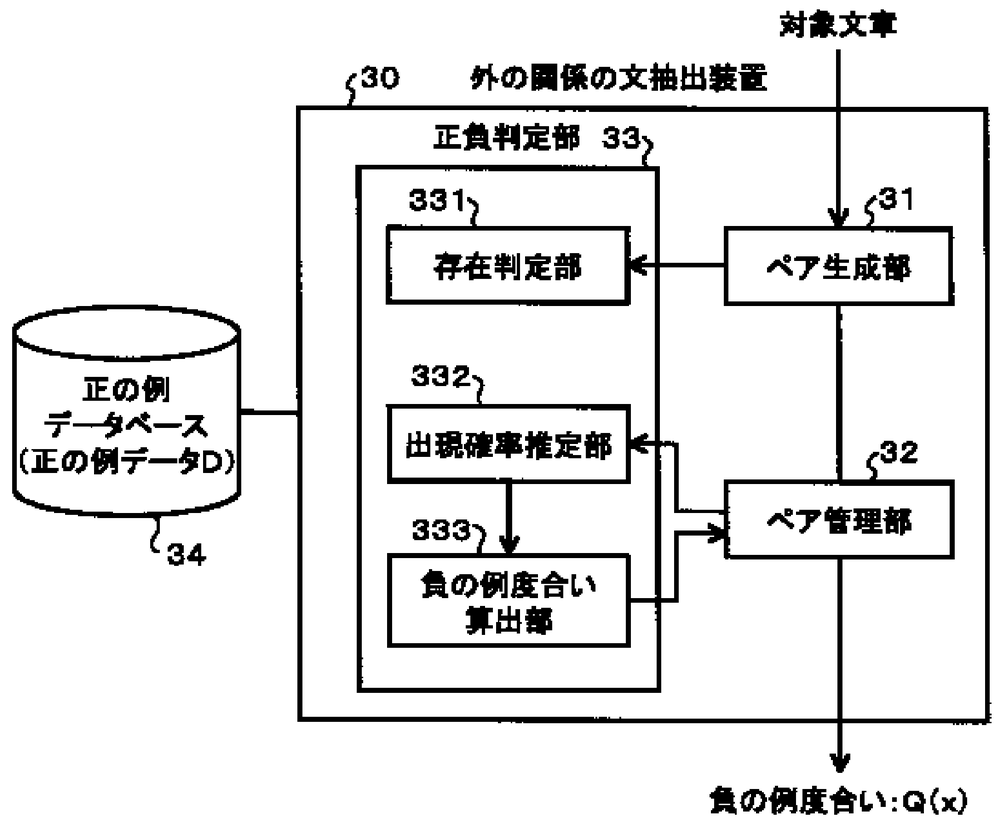
		再現率	適合率
本発明の結果			
上位	50個	3.12%	100.00%
上位	100個	5.94%	95.00%
上位	200個	11.56%	92.50%
上位	300個	16.62%	88.67%
上位	500個	25.25%	80.80%
上位	800個	34.25%	68.50%
上位	1,200個	42.44%	56.58%
上位	1,600個	48.06%	48.06%
上位	3,000個	61.38%	32.73%
上位	5,000個	70.81%	22.66%
上位	10,000個	81.12%	12.98%
上位	20,000個	87.62%	7.01%
上位	50,000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１０】

		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１１】

外の関係の文抽出装置の構成例



フロントページの続き

Ｆターム(参考)　5B009 QA14
5B064 EA18
5B091 AA15 EA02 EA04

(54)【発明の名称】　負の例予測処理方法、負の例予測処理プログラム、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、負の例予測処理を用いた日本語表記誤り検出装置、負の例予測処理を用いた外の関係の文抽出処理プログラム、負の例予測処理を用いた外の関係の文抽出装置