(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2002－91961

（Ｐ２００２－９１９６１Ａ）

(43)公開日　平成14年３月29日(2002.3.29)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ (参考)
G06F 17/27		G06F 17/27	E 5B091

審査請求　有　請求項の数４　ＯＬ　（全12頁）

(21)出願番号　特願2000－280582(Ｐ2000－280582)

(22)出願日　平成12年９月14日(2000.9.14)

特許法第30条第１項適用申請有り　2000年３月21日　社団法人情報処理学会発行の「情報処理学会研究報告　情処研報　Ｖｏｌ.2000，Ｎｏ.29」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
兵庫県神戸市西区岩岡町岩岡588－２　郵政省通信総合研究所　関西先端研究センター内

(72)発明者　内山　将夫
兵庫県神戸市西区岩岡町岩岡588－２　郵政省通信総合研究所　関西先端研究センター内

(74)代理人　100087848
弁理士　小笠原　吉義

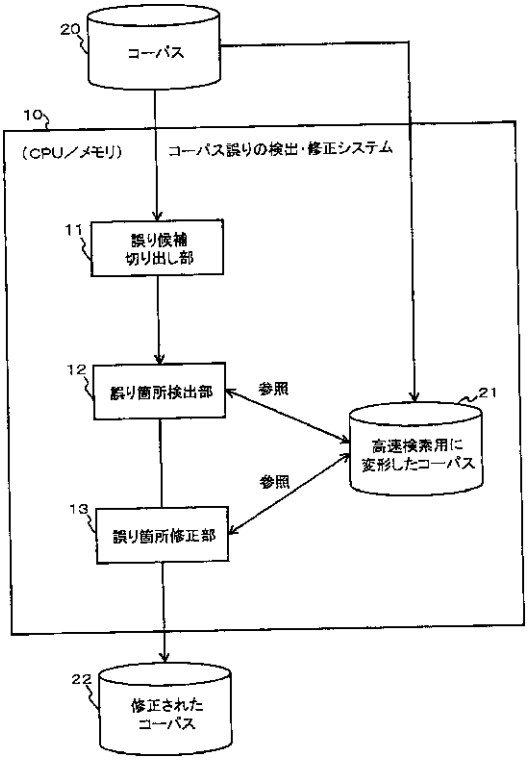
最終頁に続く

(54)【発明の名称】コーパス誤りの検出・修正システム，コーパス誤りの検出・修正処理方法およびそのプログラム記録媒体

(57)【要約】

【課題】　タグ付きコーパスにおける種々の誤りを検出し，検出した誤りを自動的に修正できるようにする。

【解決手段】　誤り候補切り出し部１１によって，タグ付きコーパスから誤り候補を切り出し，誤り箇所検出部１２によって，切り出した誤り候補のタグが誤っているかどうかを，誤り候補の正解確率，誤り候補の誤り確率および変更可能な修正候補の正解確率の算出によって評価し，誤り箇所修正部１３によって，評価結果に基づき修正候補の提示または修正されたコーパス２２を出力する。



1

【特許請求の範囲】

【請求項１】　タグ付きコーパスの誤りを検出し修正するためのコーパス誤りの検出・修正システムであって，タグ付きコーパスから誤り候補を切り出す誤り候補切り出し手段と，切り出した誤り候補のタグが誤っているかどうかを，前記誤り候補の正解確率，誤り候補の誤り確率および変更可能な修正候補の正解確率の算出によって評価する誤り箇所検出手段と，前記評価結果に基づいて修正候補の提示または修正されたコーパスを出力する誤り箇所修正手段とを備えることを特徴とするコーパス誤りの検出・修正システム。

【請求項２】　前記誤り箇所検出手段は，何個かの形態素連続における形態素情報を誤り候補として，形態素情報の誤りを検出する　ことを特徴とする請求項１記載のコーパス誤りの検出・修正システム。

【請求項３】　タグ付きコーパスの誤りを検出し修正するためのコーパス誤りの検出・修正処理方法であって，タグ付きコーパスから誤り候補を切り出す過程と，切り出した誤り候補のタグが誤っているかどうかを，前記誤り候補の正解確率，誤り候補の誤り確率および変更可能な修正候補の正解確率の算出によって評価する過程と，前記評価結果に基づいて修正候補の提示または修正されたコーパスを出力する過程とを有することを特徴とするコーパス誤りの検出・修正処理方法。

【請求項４】　コンピュータによってタグ付きコーパスの誤りを検出し修正するためのプログラムを記録した記録媒体であって，タグ付きコーパスから誤り候補を切り出す処理と，切り出した誤り候補のタグが誤っているかどうかを，前記誤り候補の正解確率，誤り候補の誤り確率および変更可能な修正候補の正解確率の算出によって評価する処理と，前記評価結果に基づいて修正候補の提示または修正されたコーパスを出力する処理とを，コンピュータに実行させるためのプログラムを記録したことを特徴とするコーパス誤りの検出・修正用プログラム記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は，計算機による言語処理システムの分野で用いられるコーパスの誤りを検出し，それを自動修正することを可能にしたコーパス誤りの検出・修正システムに関するものである。

【０００２】コーパスとは，言語分析用の電子化された言語資料である。言語処理の分野では，システムの構築にコーパスを参照することが多く，コーパスは重要な役割を果している。特に，タグ付きコーパスとは，以下のように普通の文（「車で行く。」）に特殊なタグ（品詞情報など）が付いているものをいう。

【０００３】『車（名詞）で（助詞）行く（動詞）。』

このコーパスに付けられたタグが間違っている場合もあり，このことが各研究の進捗の妨げになることも多い。

2

本発明は，このコーパス中の誤りを，決定リスト，用例ベース手法などを用いて，検出したり訂正したりするものである。

【０００４】

【従来の技術】近年，さまざまなコーパスが作られ，「教師あり機械学習」の研究をはじめとして，コーパスを用いた多種多様な研究が数多くなされている。しかし，コーパスには誤りが付きもので，この誤りが各研究の進捗を妨げる場合も多い。このため，コーパス中の誤りを検出・修正することは非常に重要なことである。

【０００５】このコーパス中の誤りを検出する試みが，最近いくつかなされ始めている。

［参考文献１］内山将夫，形態素解析結果から過分割を検出する統計的尺度，
言語処理学会誌，Vol.6 ，No.7，1999.
この参考文献１では，例えば「休憩室」という語がコーパスで「休」と「憩室」に分割されているような誤りを検出する研究について示されている。

［参考文献２］乾孝司　乾健太郎，統計的部分係り受け解析における係り受け確率の利用法--- コーパス中の構文タグ誤りの検出 ---，情報処理学会自然言語処理研究会
99-NL-134 ，1999.
この参考文献２では，コーパス中の構文的誤りを検出する研究について示されている。

【０００６】まず，上記参考文献１に記載されている技術について説明する。この参考文献１の研究では，形態素コーパスでの過分割の誤り，例えば，「休憩室」を「休」「憩室」と分割してしまう誤りを検出する方法を提案している。単語分割の問題は，情報検索において重要な問題として位置づけられている。ここでは，「分割した場合の確率」と「つなげた場合の確率」をコーパスから求め，「つなげた場合の確率」の方が圧倒的に大きい場合に，分割するのは間違いであると判定する。

【０００７】また，上記参考文献２の研究では，構文情報のコーパスでの係り先の誤りを検出する方法を提案している。コーパス中のある文節Ｘの係り先Ｙが合っているかどうかを調べる場合，コーパスからその文節Ｘがその係り先Ｙになる確率を求め，その確率が極端に小さい場合にその係り先Ｙは間違いであると判定する。

【０００８】両者の研究は，一般化して考えるとほぼ同様なことをしており，コーパスのタグが合っている確率と間違っている確率を求め，間違っている確率の方が圧倒的に大きい場合に，そのコーパスのタグを誤りとするという方法を採っている。「間違っている確率」の大きいものを間違っているものと考えるのは自然なことであり，ほとんどのコーパス修正の研究で，この種の考え方を利用することが可能であると考えられる。

【０００９】しかし，先の二つの研究で用いられた手法は，いずれも形態素の過分割，係り受け誤りと，それぞ

れその問題に特化した方法を用いて誤り検出を行っていたため，その手法の汎用性を見えにくくしている。

【００１０】参考文献１の過分割の研究では，過分割の検出に特化したような式，例えば，Ｐ（ｘ）をｘの出現率として，

Ｐ（休憩室）／（Ｐ（休）Ｐ（憩室））

が用いられている。ここで，Ｐ（休），Ｐ（憩室）の部分は，「休」「憩室」の単純な出現率を用いているが，厳密には「休憩室」という文字列が，「休」と「憩室」に分割される確率を用いるもので，近似をすでに使ったものとなっている。この近似は，データスパースネスに対処するためのものであるが，この近似自体は，過分割の検出と同じような問題でしか使えない。

【００１１】また，参考文献２の研究では，すでにできあがった構文解析システムが出す誤り確率を利用している。この構文解析システムでは，構文解析に特化した情報を数多く利用していると思われるし，また，誤りを検出する対象とするコーパス以外の情報を用いている可能性も高く，汎用的なコーパス修正とは言いにくい。

【００１２】

【発明が解決しようとする課題】上記参考文献１および参考文献２に記載されている方法では，誤り検出の適用範囲が過分割および構文的誤りというように限定されており，例えば品詞の誤りというような形態素情報の誤りを検出することができないという問題があった。また，単に誤り検出のみを対象としているため，検出した誤りをどのように修正すればよいかは人間が考えなければならず，また誤り箇所を修正するのも人間が行わなければならないという問題があった。

【００１３】本発明は上記問題点の解決を図り，タグ付きコーパスにおける種々の誤りを検出し，それに対する修正案を提示し自動修正する手段を提供することを目的とする。

【００１４】

【課題を解決するための手段】本発明は，上記課題を解決するため，タグ付きコーパスから誤り候補を切り出す誤り候補切り出し手段と，切り出した誤り候補のタグが誤っているかどうかを，誤り候補の正解確率，誤り候補の誤り確率および変更可能な修正候補の正解確率の算出によって評価する誤り箇所検出手段と，評価結果に基づいて修正候補の提示または修正されたコーパスを出力する誤り箇所修正手段とを備えることを特徴とする。

【００１５】また，本発明は，誤り箇所検出手段として，何個かの形態素連続における形態素情報を誤り候補として，形態素情報の誤りを検出する手段を持つことにより，形態素情報の誤りについても検出しその修正を実現することを特徴とする。

【００１６】本発明は，単にコーパス中の誤りを検出するだけでなく，それを修正する手段を持つことが，従来技術と大きく異なる。また，従来技術では，形態素の過分割，係り受け誤りといったそれぞれの問題に特化した誤り検出しか行われていなかったのに対し，本発明は，例えば形態素情報の誤り，構文情報の誤りといった各種の誤りを対象としてそれらを検出し，修正できる点が，従来技術と大きく異なる。

【００１７】以上の各処理手段をコンピュータによって実現するためのプログラムは，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る。

【００１８】

【発明の実施の形態】図１は，本発明のシステム構成例を示す。図中，１０は本発明に係るコーパス誤りの検出・修正システムを表す。コーパス誤りの検出・修正システム１０は，誤り候補切り出し部１１，誤り箇所検出部１２，誤り箇所修正部１３を持つ。

【００１９】図２は，図１に示すシステムの処理フローチャートである。誤り候補切り出し部１１は，修正対象であるタグ付きコーパス２０から誤り候補を何らかの単位で取り出す（ステップＳ１）。ここでは形態素情報の修正の場合，例えばコーパスから１～５個の形態素連続における形態素情報を取り出す。また，構文情報の修正の場合には，コーパスから誤り候補として，ある文節Ｘの係り先がＹのときに，他の係り先候補をＺ１，Ｚ２，Ｚ３，…としたときに，Ｘ，Ｙ，Ｚｉ（ｉ＝１，２，３，…）といった三つ組のデータを取り出す。

【００２０】次に，誤り箇所検出部１２は，誤り候補切り出し部１１が取り出した誤り候補のすべてに対して，高速検索用に変形したコーパス２１を参照して，以下の計算を行う（ステップＳ２）。

【００２１】

ａ．その誤り候補の誤り確率の算出

ｂ．そのときのシステムの確信度の算出

ｃ．そのときの修正候補の算出

なお，高速検索用に変形したコーパス２１は，誤り箇所の検出のための確率値の算出を高速化するため，コーパス２０について，データの並びの順序を変更したり，不要なデータ部分を削除したものである。例えば，形態素の生データとその品詞の組を検索するとき，形態素の生データとその品詞を連続して並べたものをオリジナルのコーパス２０からあらかじめ作っておけば，形態素の生データとその品詞の組を１個の検索キーとして，１回の検索操作で検索することができる。これが高速検索用に変形したコーパス２１である。

【００２２】その後，誤り箇所検出部１２は，取り出したすべての誤り候補のうち，確信度の高いものから，以下の処理を行う（ステップＳ３）。ただし，形態素情報の修正の場合，すでに誤り箇所もしくは正解箇所と推定された箇所を含む部分については，以下の処理は行わない。また，構文情報の修正の場合，すでに誤り箇所と推定された箇所を含む部分については，以下の処理は行わ

5

ない（形態素の場合と異なり，正解箇所と判断した箇所については，以下の処理を引き続き行う）。

【００２３】ａ．誤り候補の誤り確率が０．５以上の場合，誤り箇所と判定し，そのときの修正候補を修正候補とする。

【００２４】ｂ．誤り候補の誤り確率が０．５以上でない場合，正解箇所と判定し，その部分を以降誤り箇所とは判定しない。

【００２５】誤り箇所検出部１２は，すべての誤り候補に対して上記ステップＳ３の処理を行った後，処理を終了し，誤り箇所と判定した箇所をすべて誤り箇所と検出する（ステップＳ４）。

【００２６】誤り箇所修正部１３は，誤り箇所検出部１２が誤り箇所と検出した箇所について，ステップＳ３で修正候補としたものを修正候補として，誤りの修正候補を示す（ステップＳ５）。その結果をもとに，修正されたコーパス２２（もしくは誤り箇所指摘および修正付きコーパス）を出力する（ステップＳ６）。

【００２７】ステップＳ３では，０．５以上のものを誤り箇所と判断しているが，０．５より大きいものだけを誤り箇所と判断してもよい。以下，具体例に従ってさらに詳しく説明する。

【００２８】［コーパスの例］図３は，代表的なコーパスとしてよく知られている京大コーパスの例，図４は，そのコーパスのデータ構造の説明図である。

【００２９】京大コーパスは，図３の具体例に示すように，各文を文節に分割し，それらの係り受け関係を示すとともに，さらに各文節を形態素に分割して各形態素の品詞その他の詳細な情報を持たせたものである。

【００３０】おおよそ一文が図３に示すように構成され，一文が終わるとＥＯＳ(end of sentence）の記号が付与される。すなわち，図４（Ａ）のように，＃から始まりＥＯＳで終わる部分が一文に関する情報である。一文に関する情報は，図４（Ｂ）に示すように，＊から始まる文節に関する情報によって構成される。

【００３１】文節に関する情報の部分には，図４（Ｃ）のように，＊に続く第１カラム目に何番目の文節であるかを示す文節番号が記述され，第２カラム目には，その文節の係り先の文節番号が記述される。第２カラムの数字の次に続くアルファベットは，Ｄが通常の係りを表し，Ｐ，Ｉの場合には並列的な係り，Ａの場合には同格的な係りを意味する。続くＥＯＳ，＃，＊以外のものから始まる行は，形態素情報を表している。

【００３２】形態素情報の部分には，図４（Ｄ）のように，第１カラムに生データで出現したままの形の形態素が記述され，第２カラムに読みの情報が記述され，第３カラムに変化する形態素の場合は基本形を，そうでない場合は＊が記述される。また，第４カラムに品詞が記述され，第５カラムに品詞細分類が記述され，第６カラムに変化する形態素の場合は活用型を，そうでない場合は

6

＊が記述される。第７カラムにも形態素の活用形に関する情報が記述される。

【００３３】例えば，第３図に示す２行目の「* 0 26D 」は，第０番目（先頭）の文節を表し，この文節の係り先が第２６番目の文節（「示した」）であることを意味している。また，３行目の「村山　むらやま　＊　名詞　人名　＊　＊」は，生データの形態素が「村山」，その読みが「むらやま」，変化しないので第３カラムが「＊」，品詞は「名詞」，品詞細分類は「人名」，変化する活用型ではないので，続くカラムは「＊」，「＊」となっている。

【００３４】［コーパス修正のための評価式］コーパスの修正の課題は，このタグは正解，また，このタグは誤りというものがふられたデータがないため，基本的に「教師なし学習」の問題となる。このため，コーパス修正には何らかの基準が必要となる。先に述べた参考文献１，２の二つの先行研究では，以下の評価基準を利用して，コーパス誤り検出の教師なし問題を解いていることになっている。

〔参考文献１の方法〕

評価式＝（分割しない場合の出現率）／（分割した場合の出現率）

これを強いて一般化して表すと，次のように表すことができる。

【００３５】評価式＝（修正後のタグが正しい確率）／（修正前のタグが正しい確率）

〔参考文献２の方法〕

評価式＝（修正前のタグが誤っている確率）

これらの評価式の値が大きい場合には，タグが誤っている可能性が高いとする。クラスが二つしかない問題の場合には，上記の二つの基準は等価となる。しかし，これらの評価式は，主として誤りの検出を考慮したものになっており，検出したコーパス誤りをどのように修正するのがよいかを考慮したものにはなっていない。

【００３６】本発明では，コーパス誤りを検出する評価式として，

評価式＝修正前のタグが誤っている確率

を用い，それを修正するための評価式として，

評価式＝修正後のタグが誤っている確率

を用いることにより，コーパス誤りの自動修正を可能にする。

【００３７】［確率値の算出方法］「修正前のタグが誤っている確率」や「修正後のタグが正しい確率」といっても，これをどのようにして簡単に求めるかが次の課題となる。ここでは，まず「修正前のタグが誤っている確率」の算出方法を，具体的な処理の例に従って説明する。

【００３８】図５は，決定リストを用いる場合の確率値算出の処理の流れを示す。まず，ステップＳ１０では，誤り候補について変更可能な候補をコーパスから取り出

す。次にステップＳ１１では，何種類かのパターンを定義し，そのパターンごとに，以下の計算を行う。
ａ．誤り候補の正解確率の算出
今のパターンの形でのコーパスでの誤り候補の総出現数を，今のパターンの総出現数で割ったものを誤り候補の正解確率とする。
ｂ．誤り候補の誤り確率の算出
１から誤り候補の正解確率を引いたものを誤り候補の誤り確率とする。
ｃ．変更可能な候補ｉの正解確率の算出
今のパターンの形でのコーパスでの変更可能な候補ｉの総出現数を，今のパターンの総出現数で割ったものを変更可能な候補ｉの正解確率とする。
ｄ．修正候補の算出
ｃで計算した変更可能な候補のうち，最も正解確率が大きいものを修正候補とする。
ｅ．このときのシステムの確信度の算出
誤り候補の正解確率と，ｄで選んだ修正候補の正解確率の大きい方をこのときのシステムの確信度とする。
【００３９】次にステップＳ１２では，ステップＳ１１で求めた全パターンのうち，最も確信度の大きいパターンのときの誤り候補の誤り確率，修正候補，確信度を，その誤り候補の誤り確率，修正候補，確信度とする。
【００４０】なお，この例では，ｅの確信度として，誤り候補の正解確率と，ｄで選んだ修正候補の正解確率の大きい方を用いるとしているが，ａ，ｂで求めた誤り候補の正解確率と誤り確率の大きい方を用いることにしてもよい。
【００４１】構文情報の修正の場合には，確率値算出の処理が上記の処理と少々変わっているので，図６にその処理の流れを示す。
【００４２】誤り候補としては，ある文節Ｘの係り先がＹのときに，他の係り先候補をＺ１，Ｚ２，Ｚ３，…として，Ｘ，Ｙ，Ｚｉ（ｉ＝１，２，３，…）といった三つ組のデータが誤り候補の単位として，取り出されている。そこで，この状況下で以下の計算を行う。まず，ステップＳ２０では，変更可能な候補としてはＺｉを用いる。
【００４３】次にステップＳ２１では，何種類かのパターンを定義し，そのパターンごとに，以下の計算を行う。
ａ．誤り候補の正解確率の算出
今のパターンの形でのコーパスでの，Ｙが係り先になる総数を，今のパターンの総数で割ったものを誤り候補の正解確率とする。
ｂ．誤り候補の誤り確率の算出
１から誤り候補の正解確率を引いたものを誤り候補の誤り確率とする。
ｃ．変更可能な候補ｉの正解確率の算出
今のパターンの形でのコーパスでの，Ｚｉが係り先になる総数を，今のパターンの総数で割ったものを変更可能な候補Ｚｉの正解確率とする。
ｄ．修正候補の算出
Ｚｉを修正可能な候補とする。
ｅ．このときのシステムの確信度の算出
誤り候補の正解確率と，修正候補Ｚｉの正解確率の大きい方をこのときのシステムの確信度とする。
【００４４】ステップＳ２２では，ステップＳ２１で求めた全パターンのうち，最も確信度の大きいパターンのときの誤り候補の誤り確率，修正候補，確信度をその誤り候補の誤り確率，修正候補，確信度とする。
【００４５】京大コーパスについての確率値算出の具体例を説明する。京大コーパスについて，読点「、」の形態素情報の統計をとってみると，図７（Ａ）に示すような結果が得られる。この統計情報は，ちょっと見ただけでも「特殊　読点」となっているデータが圧倒的に大きく，他は誤っているということが予想される。ここで２行目の「、　、　、　＊　＊」の誤りの確率を考えてみる。
【００４６】まず，これの正解確率は，その出現数を総数で割ったものと考えてよい。
【００４７】
正解確率＝３／（２６５４０＋３＋２＋１）
一方，誤り確率は１から正解確率を引いたものと考えられるので，
誤り確率＝１－３／（２６５４０＋３＋２＋１）
となる。そこで，本実施の形態では，誤り確率の求め方として，基本的にこの方法を用いる。
【００４８】しかし，単にこれだけでは確率の求め方として粗すぎる場合がある。京大コーパスについて，例えば「の」の形態素情報の統計をとってみると，図７（Ｂ）のような結果が得られる。ここで，頻度が１９１の「の　の　だ　判定詞　＊　判定詞　ダ列特殊連体形」の誤り確率を求めると，
誤り確率＝１－１９１／（２５７３９＋１６０１＋…）
＝９９．３％
となって，ほとんど誤っていると判定される。「の　の　だ　判定詞　＊　判定詞　ダ列特殊連体形」が正しい場合も数多くあり，この単純な方法では，正しいのにこれを全部誤っていると推定してしまう。
【００４９】そこで，本実施の形態では，確率値の算出に用例ベース手法や決定リスト手法を利用する。用例ベース手法の参考文献としては，以下の参考文献３があり，決定リスト手法の参考文献としては，以下の参考文献４，５がある。
［参考文献３］村田真樹，内元清貴，馬青，井佐原均，排反な規則を用いた文節まとめあげ，情報処理学会論文誌，(2000).
［参考文献４］David Yarowsky, Decision lists for lexical ambiguity resolution :Application to accent

9

restoration in Spanish and French, 32th Annual Meetingof the Associtation of the Computational Linguistics,(1994), pp.88-95.

［参考文献５］新納浩幸，複合語からの証拠に重みをつけた決定リストによる同音異義語判別，情報処理学会論文誌，Vol.39，No.12,(1998).

用例ベース手法は，いま解きたいものと良く似た用例を集め，その用例集合での出現率を確率値とする手法である。

【００５０】「のような」の場合，「の」は８４個あってすべて「の　の　だ　判定詞　＊判定詞　ダ列特殊連体形」であるので，正解確率１００％，誤り確率０％となり，これを間違って誤りと検出することがなくなる。用例ベース的確率算出方法は，バックオフによる確率推定を極端なまで行ったことに相当する。また，誤り修正の場合，自分自身だけの事例を用いると一つも誤りを検出できなくなるので，最低自分以外に一つ，合計二つ以上の事例をもってくる必要がある。

【００５１】一方，決定リスト手法は，多くの素性に展開し各素性の確信度を求め，確信度の最も高い素性（パターン）のときの，正解確率と誤り確率を用いる方法である。前述した「の」の例の場合，「の」「のような」「名詞＋の」「の＋助動詞」などと，いろいろなパターンでの確率を求める（ただし，総数が１の素性は用いない）。この結果を京大コーパスを用いて計算すると，図８のようになる。

【００５２】図８における「判定詞の場合の数」は，京大コーパスで各素性に適合する事例における「の」が判定詞の場合の数で，「総数」は京大コーパスで各素性に適合する事例の総数である。例えば，「のような」のパターンは，判定詞の「の」だけが８４個出現したことを意味し，「の＋助動詞」のパターンでは，判定詞の「の」が１８７個，それ以外の「の」が１個出現したことを意味する。

【００５３】このデータからの正解確率，誤り確率の求め方は，先に述べた例と同じで，

正解確率＝１８７／１８８

誤り確率＝１－（１８７／１８８）

などの計算をして求める。

【００５４】また，「確信度」はその規則の確らしさを意味するものであり，この確信度としては，正解確率と誤り確率のうち大きい方の値を用いる。例えば，１行目の「のような」は，確信度１００％でほぼ正しい情報と推測されることになる。この規則は，上記参考文献３でいう排反な規則に相当する。

【００５５】決定リストでは，この図８の最上位にある，この規則を用いることになり，誤り確率は０となって，用例ベースと同じく「のような」の「の」は，判定詞で正しいと推定され，間違って誤りと推定することはない。図８の上の二行の情報がないときは，誤り確率９

10

９．３％，確信度９９．３％で誤っていると判定される。

【００５６】次に「修正後のタグが正しい確率」の求め方であるが，これは，図７（Ａ）の読点の簡単な場合で考えると，「修正後のタグ」は頻度の最も大きい「、　、＊　特殊　読点　＊　＊」とすればよく，これが正しい確率は，これの出現数を総数で割ったもの，すなわち，９９．９９％（＝２６５４０／２６５４３）となる。

【００５７】以上は単純な場合の例であるが，用例ベース手法，決定リスト手法の場合ともに，誤り確率などを求めた事例集合でこの計算をして，「修正後のタグが正しい確率」を求めればよい。

【００５８】もちろん確率値を算出する方法は，用例ベース手法，決定リスト手法に限られるわけではなく，例えば最大エントロピー法など，その他の手法を用いて確率値を求めることもできる。

【００５９】［形態素情報の修正例］以下では，形態素情報のコーパス修正を試みた結果について述べる。まず，対象とする京大コーパスでの形態素情報の調査を行った。この結果を図９に示す。図９における全形態素数はコーパスにあったすべての形態素の数を意味する。また，曖昧形態素数はコーパスにあった形態素のうち，コーパス中の他の形態素と表記が同じであった形態素の数を意味する。例えば「の　の　＊　助詞　格助詞＊　＊」，「の　の　＊　助詞　接続助詞　＊　＊」といったものは，表記が同じ「の」で異なる形態素なので曖昧形態素と考える。

【００６０】また，この調査では，５つまでの形態素連続までは「では」と「で｜は」のように形態素の区切りが異なるものが他にある場合も曖昧形態素と考えている（つまり，この場合，「では」「で」「は」はそれぞれ曖昧形態素となる）。

【００６１】図９中の「読み情報あり」と「読み情報なし」は，京大コーパスが読み情報に弱いという理由から設定したもので，「読み情報あり」は，読み情報も含めて曖昧形態素の数を数えたもので，「読み情報なし」は，読み情報を省いて曖昧形態素の数を数えたものを意味する。全形態素数は「読み情報あり」と「読み情報なし」とで変わることはない。

【００６２】例えば「読み情報なし」では，「日　ひ　＊　名詞　時相名詞　＊　＊」と「日　び　＊　名詞　時相名詞　＊　＊」のように読み情報のみが異なる場合，これらを異なる形態素として扱わない。

【００６３】図９からわかるように，京大コーパス約２万文には，４８７，６９１形態素が存在しており，人手で５０万の形態素を徹底的に調べあげるとコーパス修正ができるがそれは非常に大変である。また，曖昧形態素数は，読み情報の修正を諦めたとしても，２７０，５３４形態素存在しており，修正範囲を曖昧な形態素に絞っ

たところで網羅的に人手で修正するのは困難である。

【００６４】曖昧形態素数の異なりは，５，５３９であるので，曖昧形態素の種類ごとにまとめて出力させ，それを見て人手で修正することも可能かとも思われるが，各種類ごとに多数の事例が出力されると思われ，それを用いた修正も若干無理があると思われる。

【００６５】以上のことから，コーパス修正は難しい問題であることがわかる。このため，このコーパス修正を容易に行う技術を確立することは重要である。

【００６６】以下に述べる形態素情報の修正の例では，読み情報は対象から外している。そこで，図１の高速検索用に変形したコーパス２１では，入力したコーパス２０を変形し，読み情報の項目を消している。「タグが誤っている確率」の算出には，前に述べたように用例ベース手法と決定リスト手法とを利用する。

【００６７】まず，１～５個の形態素連続における形態素情報を誤りの候補とする。この誤りの各候補に対し，「タグが誤っている確率」と「確信度」と「修正後のタグ」を算出する。次に，確信度の大きい誤り候補から順に欲張り法でコーパスを修正する。このとき，各修正箇所には先に算出した「タグが誤っている確率」と「修正後のタグ」を付与しておく。「タグが誤っている確率」が０．５より大きい形態素のタグが誤っているものと判定され，「修正後のタグ」に修正される。０．５以下の形態素のタグは正しいものと判断され，修正の対象とならない。

【００６８】「タグが誤っている確率」と「確信度」と「修正後のタグ」の算出方法は，以下のとおりである。まず，誤り候補から変更可能な候補をコーパスより取り出す。ここで，変更可能な候補とは，表記が同じものである。例えば「ロシア　＊名詞　普通名詞　＊　＊」が誤り候補の場合，「ロシア　＊　名詞　地名　＊＊」が変更可能な候補として取り出される。

【００６９】ここで，用例ベース手法の場合には，誤り候補のまわりの形態素の状態が最もよく似ている用例を集め，その用例集合で前述した方法で「タグが誤っている確率」と「修正後のタグ」を推定する。最もよく似ている用例の集め方は，候補の形態素から出発し，それに対して，その前後の形態素の品詞，品詞細分類，残りの全情報を順次追加していき，さらにその隣の形態素からもそのような情報を順次追加する。これを繰り返し，検出される用例が１個だけになる直前の状態のときの用例を利用する。

【００７０】「確信度」は，ここでは「タグが誤っている確率」と「タグが正しい確率」のうち大きい方の値としている。「確信度」を図５に示した例のように，誤り候補の正解確率と，修正候補の正解確率の大きい方の値としてもよい。

【００７１】また，決定リスト手法の場合には，以下で説明する１６個の素性を用いて，前述した方法を用いて「タグが誤っている確率」と「確信度」と「修正後のタグ」を推定する。１６個の素性については，まず，各形態素の情報として以下の四つのパターンの情報を考え，この四つのパターン情報を，候補となっている形態素の前後二つの形態素についてあらゆる組合わせを作って，合計１６個の素性を作り，それを決定リスト用の素性とする。

（１）情報なし

（２）品詞情報のみ

（３）品詞情報と品詞細分類情報のみ（活用する形態素の場合には，品詞情報と活用形のみを用いる）

（４）形態素情報すべて

上記の方法でコーパス修正を行った結果は，以下のとおりであった。

【００７２】用例ベース手法では，５９１個がタグ誤りと検出され，決定リスト手法では，４，０５４個がタグ誤りと検出された。その検出されたデータの精度を，図１０に示す。

【００７３】図１０中の「ランダム３００個」は，「誤り確率」のことを考慮せずにコーパスの先頭３００個を調査したときの精度で，ほぼ平均精度に相当する。「上位ｘ個」は，集計したデータを「誤り確率」に基づいてソートし，「誤り確率」の上位ｘ個のものの精度を調べたものである。「検出精度」は，誤り部分を正しく検出した箇所の数を総数で割ったもので，「修正精度」は，誤り部分を正しく修正した箇所の数を総数で割ったものである。また図１０中の「不明」は，正否がはっきりしない場合の数である。「不明」としたものには，副詞と名詞，サ変名詞と普通名詞，普通名詞と動詞連用形など，タグの定義のゆれに関係しそうなものも含めている。検出精度，修正精度の算出では，検出，修正を失敗したものとして扱っている。

【００７４】今回の実験では，図１０のように，用例ベース手法よりも決定リスト手法の方が抽出数，抽出精度ともによかった。ただし，この結果は本実施の形態における素性の設定状況によるかもしれず，常に決定リスト手法の方がよいとは限らない。

【００７５】決定リスト手法では，抽出総数が約４，０００で平均精度（図１０の「ランダム３００個」）が５０%程度あるので，おおよそこの４，０００のデータを見るだけで２，０００個の誤りを修正できる計算となる。また，上位での精度は７０%～８０%と比較的高く誤りを検出できており，この精度ならば人手でこれをチェックしつつコーパス修正をするのもそれほど負担にならないと思われ，十分実用的にコーパス修正に利用可能であると考えられる。

【００７６】図１１は，決定リスト手法の上位での修正結果の例を示している。該当箇所の欄に×印をつけているものは誤り検出失敗を意味する。検出の上位には，「、、　＊　＊」といったコーパス作成中に何らか

のデータ作成ミスが生じたのではないかと思われる明らかな誤りも含まれている。

【００７７】「の　＊　連体詞　＊　＊　＊」「は　＊　助詞　格助詞　＊　＊」というアノテーターによるミスと思われる誤りもある。「～ぐらいの～」を誤ってコーパス誤りと推定しているが，これはコーパス中の他の誤りが原因となっている。「～ぐらいの～」の「の」はほとんど判定詞「だ」であるが，コーパスで格助詞「の」としている箇所が二つあるため，決定リストの一つの素性「～ぐらいの～」における判定詞「だ」のタグが合っている確率が１にならず，誤りと検出してしまっている。

【００７８】決定リスト手法の場合には，手法の原理が簡単であるために，誤り検出を失敗したとき，それならこっちのほうが誤っているのではないかと推測することが容易なので，誤り検出を失敗したとしても，副産物として他の誤りを検出できる可能性が高い。

【００７９】［構文情報の修正例］次に，構文情報の修正結果について述べる。本実験では，京大コーパスのうち，１９９５年1月１０日までの約１万文のデータを利用した。以下で修正方法を述べる。ある文節Ｘの係り先がＹのときに，その文節Ｘの係り先のタグが正しいかどうかを判定する場合，他の係り先候補をＺ１，Ｚ２，Ｚ３，…としたとき，Ｘ，Ｙ，Ｚｉ （ｉ＝１，２，３，…）の三つ組のデータに対し，ＹとＺｉ の比較でＹが係り先となる確率とＺｉ が係り先になる確率を求め（この二つの確率の求め方は後で述べる），これらの確率の大きい方を「確信度」とし，Ｚｉ が係り先になる確率を「誤っている確率」とし，Ｚｉ を「修正タグ」とする。

【００８０】これをすべてのＺ１，Ｚ２，Ｚ３，…に対して計算し，このうち，「誤っている確率」が最も大きいＺｉの「誤っている確率」と「修正タグ」を，文節Ｘに付与する。「誤っている確率」が０．５よりも大きい文節の係り先タグは誤っていると判断し，その係り先タグは「修正タグ」に修正する。

【００８１】次に，Ｘ，Ｙ，Ｚｉ の三つ組のデータにおいて，Ｙが係り先となる確率とＺiが係り先になる確率の求め方を記述する。この確率の算出には，決定リストを利用する。文節情報のＡパターンとして以下を定義する。

（１）情報なし
（２）付属語の品詞の情報
（３）付属語の品詞と品詞細分類の情報
（４）付属語の品詞と品詞細分類の情報と，自立語の品詞
（５）付属語の品詞と品詞細分類の情報と，自立語の品詞と分類語彙表の分類番号の上位５桁
（６）付属語の品詞と品詞細分類の情報と，自立語の品詞と分類語彙表の分類番号の上位５桁と単語自体

また，文節情報のＢパターンとして以下を定義する。

（１）情報なし
（２）自立語の品詞
（３）自立語の品詞と品詞細分類
（４）自立語の品詞と品詞細分類と分類語彙表の分類番号の上位５桁
（５）自立語の品詞と品詞細分類と分類語彙表の分類番号の上位５桁と単語自体

文節ＸにはＡパターンを，文節Ｙ，Ｚｉ にはＢパターンを利用し，すべての各パターンの組合わせ，つまり，６×５×５の素性を作る。また，ＹとＺｉ は，どちらが文で先に出現しているかも素性とし，合計６×５×５×２の素性をこの決定リストの素性とする。

【００８２】この素性ごとに，コーパスより文節Ｙが係り先になる場合の数と，Ｚｉ が係り先になる場合の数を求め，それぞれをその和で割ることでそれぞれの確率値を求める。

【００８３】また，このとき大きい方の確率値を確信度とする。この計算をすべての素性で行ってやり，確信度が最も大きいときの素性の，Ｙが係り先となる確率とＺｉ が係り先になる確率を，Ｘ，Ｙ，Ｚｉ の三つ組のデータにおけるその確率とする。ただし，文節Ｙが係り先になる場合の数が１で，そうでない場合の数が０となる素性のデータは削除する。

【００８４】この方法で実験を行った結果を，図１２に示す。また，正しく構文誤りを修正できたものの例を，図１３に示す。図１３において，墨付き括弧の記号で囲まれている文節の係り先が，コーパスでは一重下線の文節であったが，二重下線の文節に正しく修正できたことを示している。図１２のように抽出数がおよそ１，４５６で，平均検出精度が１３％なので，この１，４５６のデータから２００個くらい誤りを検出できると期待される。精度が格段に高いと言えないがそれなりにコーパスの誤り修正ができており，本手法の汎用性の検証には十分であると思われる。

【００８５】

【発明の効果】以上説明したように，本発明によれば，単にコーパス誤りの指摘だけでなく，誤った部分をどう直せば良いかも示すため，コーパス修正の効率が向上する。

【図面の簡単な説明】

【図１】本発明のシステム構成例を示す図である。
【図２】本システムの処理フローチャートである。
【図３】京大コーパスの例を示す図である。
【図４】京大コーパスのデータ構造の説明図である。
【図５】決定リストを用いる場合の確率値算出の処理の流れを示す図である。
【図６】決定リストを用いて構文情報を修正する場合の確率値算出の処理の流れを示す図である。
【図７】形態素情報の統計情報を示す図である。
【図８】決定リストによる確率値算出方法の例を示す図

15

である。

【図９】形態素情報の調査結果の例を示す図である。

【図１０】形態素情報の修正結果の例を示す図である。

【図１１】形態素誤り修正結果の例を示す図である。

【図１２】構文情報の修正結果の例を示す図である。

【図１３】正しく構文誤りを修正できた例を示す図である。

【符号の説明】

16

１０　コーパス誤りの検出・修正システム

１１　誤り候補切り出し部

１２　誤り箇所検出部

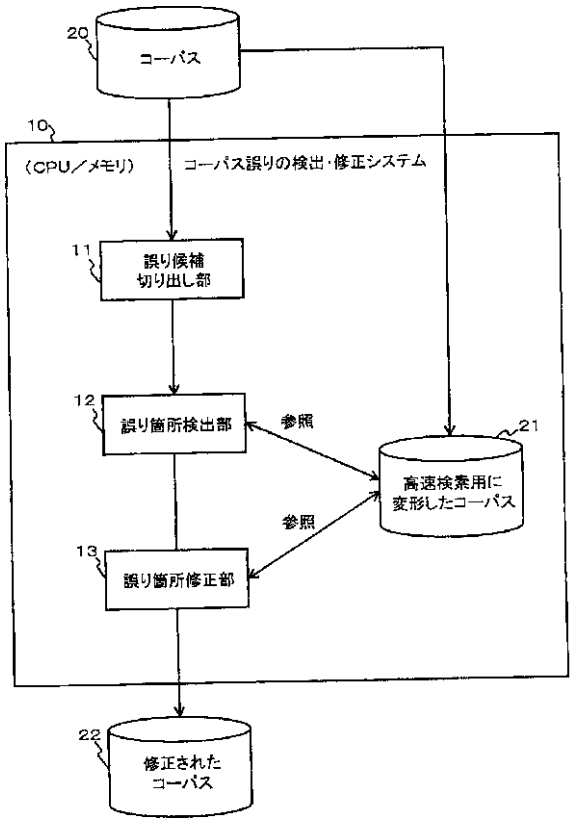
１３　誤り箇所修正部

２０　コーパス

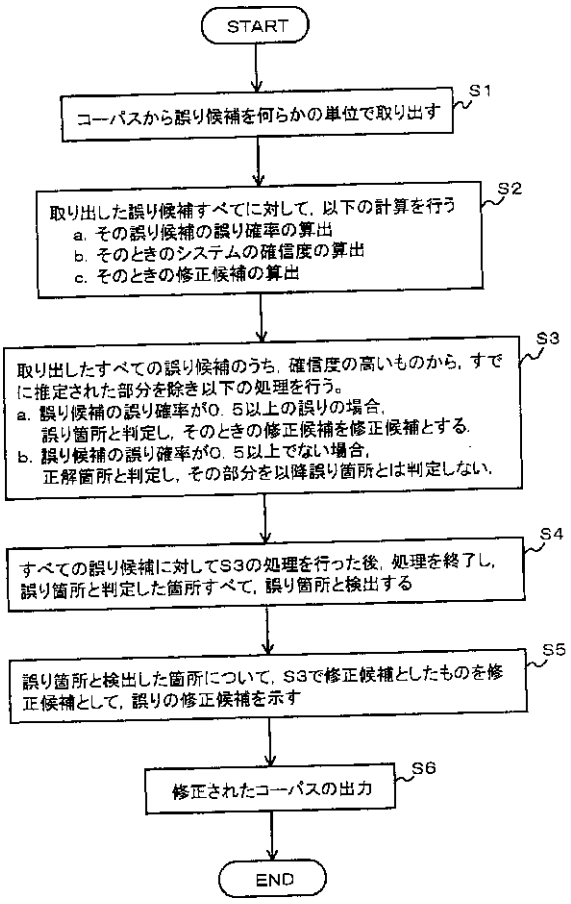
２１　高速検索用に変形したコーパス

２２　修正されたコーパス

【図１】



【図２】



【図７】

(A)　「、」の形態素情報の統計

、 、 * 特殊 読点 * *	26,540
、 、 、 * *	3
、 未定義語 * 名詞 サ変名詞 * *	2
、 、 * 特殊 記号 * *	1

(B)　「の」の形態素情報の統計

の の * 助詞 接続助詞 * *	25,739
の の * 助詞 格助詞 * *	1,601
の の * 名詞 形式名詞 * *	1,350
の の のだ 助動詞 * ナ形容詞 語幹	398
の の だ 判定詞 * 判定詞 ダ列特殊連体形	191
の の の * *	1
の の * 名詞 普通名詞 * *	1
の の * 連体詞 * * *	1

【図８】

「名詞＋のような」の場合の決定リストによる確率値の算出方法

各素性	確信度	正解確率	誤り確率	判定詞の場合の数	総数
のような	100%	100%	0%	84 個	84 個
の＋助動詞	99.5%	99.5%	0.5%	187 個	188 個
の	99.3%	0.7%	99.3%	191 個	29,282 個
名詞＋の	99.2%	0.8%	99.2%	162 個	20,220 個
.....					

【図３】

京大コーパスの例

```
# S-ID:950101003-001 KNP:96/10/27MOD:
* 0 26D
村山 むらやま * 名詞 人名 * *
富市 とみいち * 名詞 人名 * *
首相 しゅしょう * 名詞 普通名詞 * *
は は * 助詞 副助詞 * *
* 1 2D
年頭 ねんとう * 名詞 普通名詞 * *
に に * 助詞 格助詞 * *
* 2 6D
あたり あたり あたる 動詞 * 子音動詞 ラ行基本連用形
* 3 6D
首相 しゅしょう * 名詞 普通名詞 * *
官邸 かんてい * 名詞 普通名詞 * *
で で * 助詞 格助詞 * *
* 4 6D
内閣 ないかく * 名詞 普通名詞 * *
記者 きしゃ * 名詞 普通名詞 * *
会 かい * 名詞 普通名詞 * *
と と * 助詞 格助詞 * *
   .
   .
   .
* 24 25D
至らない いたらない 至らない 形容詞 * イ形容詞 アウオ段基本形
と と * 助詞 格助詞 * *
の の * 助詞 接続助詞 * *
* 25 26D
見通し みとおし * 名詞 普通名詞 * *
を を * 助詞 格助詞 * *
* 26 -1D
示した しめした 示す 動詞 * 子音動詞 サ行タ形
。 。 * 特殊 句点 * *
EOS
```

【図４】

コーパスのデータ構造の例

(A) 文情報

#	(一文に関する情報)	EOS

(B) 一文に関する情報

*	(文節に関する情報)	*	(文節に関する情報)	*	…	EOS

(C) 文節に関する情報

(*)	文節番号	係り先の文節(D/P/I/A)	形態素情報

(D) 形態素情報

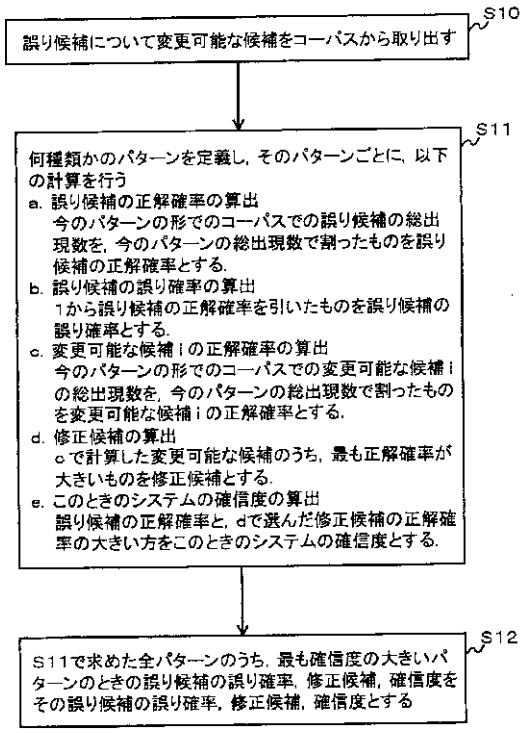
生データ	読み	基本形	品詞	品詞細分類	活用形	活用形

(E) 形態素情報の例

村山	むらやま	*	名詞	人名	*	*
示した	しめした	示す	動詞	*	子音動詞	サ行タ形

【図５】

確率値算出の処理の流れ(決定リストの場合)



【図９】

形態素情報

	読み情報あり	読み情報なし
全形態素数	487,691	487,691
曖昧形態素数 (のべ)	275,291	270,534
曖昧形態素数 (異なり)	5,957	5,539

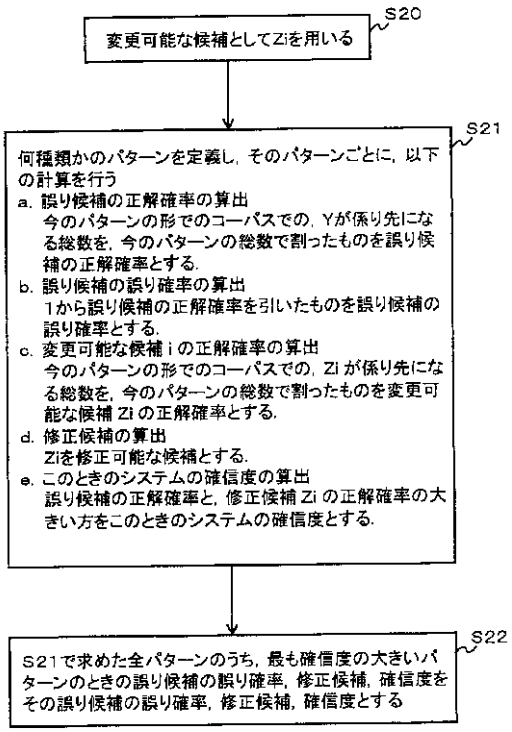
【図１２】

形態素情報の修正結果

	決定リスト手法 (抽出総数 1,456 個)		
	検出精度	修正精度	不明
ランダム 300 個	13%(40/300)	12%(37/300)	9 個
上位 50 個	26% (13/50)	20% (10/50)	2 個
上位 100 個	28%(28/100)	24%(24/100)	2 個
上位 150 個	25%(38/150)	23%(34/150)	2 個
上位 200 個	25%(50/200)	23%(45/200)	5 個
上位 250 個	24%(61/250)	22%(56/250)	5 個
上位 300 個	23%(69/300)	21%(64/300)	6 個

【図６】

確率値算出の処理の流れ（決定リストの場合，構文情報の修正の場合）



【図１０】

構 文 情 報 の 修 正 結 果

	用例ベース手法（抽出総数 591 個）			決定リスト手法（抽出総数 4,054 個）		
	検出精度	修正精度	不明	検出精度	修正精度	不明
ランダム 300 個	43%(130/300)	41%(123/300)	38 個	53%(160/300)	49%(146/300)	17 個
上位 50 個	60% (30/50)	58% (29/50)	0 個	82% (41/50)	78% (39/50)	0 個
上位 100 個	55% (55/100)	52% (52/100)	6 個	70% (70/100)	67% (67/100)	0 個
上位 150 個	49% (74/150)	47% (70/150)	9 個	71%(107/150)	66% (99/150)	2 個
上位 200 個	46% (91/200)	44% (87/200)	17 個	74%(147/200)	69%(137/200)	2 個
上位 250 個	44%(110/250)	41%(103/250)	23 個	78%(194/250)	73%(183/250)	3 個
上位 300 個	42%(125/300)	39%(117/300)	31 個	76%(229/300)	71%(212/300)	4 個

【図１１】

形態素誤り修正結果（上位２０個，誤り確率1.0000～0.9998）

該当箇所	修正前	修正後
けいはんなの経営など、	の＊名詞 形式名詞＊＊	の＊助詞 接続助詞＊＊
3日夜、札幌発関西	、＊特殊 記号＊＊	、＊特殊 読点＊＊
米朝交渉の駆け引きに悪い	ののの＊＊	の＊助詞 接続助詞＊＊
さくらももこの原作で、	の＊連体詞＊＊＊	の＊助詞 接続助詞＊＊
では南方の鉄鉱石の開発	の＊名詞 普通名詞＊＊	の＊助詞 接続助詞＊＊
中西部、少数民族地区	、、＊＊	、＊特殊 読点＊＊
一年ぐらいの期間をかけて×	のだ 判定詞＊判定詞 ダ列特殊連体形	助詞 接続助詞 の＊助詞 接続助詞＊＊接続助詞 助詞
区の同、坂本舘子さん	、＊名詞 サ変名詞＊＊	、＊特殊 読点＊＊
年にドルと金との	と＊名詞 普通名詞＊＊	と＊助詞 格助詞＊＊
をとっても、極めて重要である。	、、＊＊	、＊特殊 読点＊＊
当時としては壮大なスケールの	は＊助詞 格助詞＊＊	は＊助詞 副助詞＊＊
二日夜、多数のロシア	、、＊＊	、＊特殊 読点＊＊
鈴木いどむが左足で	が＊助詞 接続助詞＊＊	が＊助詞 格助詞＊＊
にそうだった。	。。＊＊	。＊特殊 句点＊＊
とみたい。	。。＊＊	特殊 句点 。＊特殊 句点＊＊句点 特殊
などをのどに詰まらせる事故	にだ 形容詞＊ナ形容詞 ダ列基本連用形	に＊助詞 格助詞＊＊
のソフトだ。	。。＊＊	。＊特殊 句点＊＊
自転車かごにほうり込んでいく	にる 動詞＊母音動詞 基本連用形	に＊助詞 格助詞＊＊
については、	、＊名詞 サ変名詞＊＊	、＊特殊 読点＊＊
「ジュラシック・パーク」ぐらいのもの×	のだ 判定詞＊判定詞 ダ列特殊連体形	の＊助詞 接続助詞＊＊接続助詞 助詞

【図１３】

正しく構文誤りを修正できたと判断したもの

子供【二人の】家族四人が困ることのないようにとお願いした

【その】建設資金、千二百万カナダドルは、すべてチャリティー活動や寄付で賄い「大半が香港マネーと言ってよい」と古舘集局長。

それ以後の名人に勝局への「会心の一手・一局」をピックアップしてもらい、【その】当時の状況を再現してみた。

当たり前のように【存在し、】軽視されてきた水が、いま復権を叫んでいる。

男女間の【賃金や】教育水準の格差などを加味すると、一挙に１７位まで低下することが示されている。

フロントページの続き

(72)発明者　内元　清貴
兵庫県神戸市西区岩岡町岩岡588－２　郵政省通信総合研究所　関西先端研究センター内

(72)発明者　馬　青
兵庫県神戸市西区岩岡町岩岡588－２　郵政省通信総合研究所　関西先端研究センター内

(72)発明者　井佐原　均
兵庫県神戸市西区岩岡町岩岡588－２　郵政省通信研合研究所　関西先端研究センター内

Ｆターム(参考)　5B091 EA04