

(11)特許出願公開番号

特開2010-66957

(P2010-66957A)

(43) 公開日 平成22年3月25日(2010.3.25)

(51) Int.Cl.
G06F 17/21

F I
G O 6 F 17/21 5 5 0 Z

テーマコード (参考)
5B109

審査請求 未請求 請求項の数 10 O L (全 15 頁)

(21) 出願番号	特願2008-231956 (P2008-231956)	(71) 出願人	000004226
(22) 出願日	平成20年9月10日 (2008. 9. 10)		日本電信電話株式会社
			東京都千代田区大手町二丁目3番1号
		(74) 代理人	100121706
			弁理士 中尾 直樹
		(74) 代理人	100066153
			弁理士 草野 卓
		(74) 代理人	100128705
			弁理士 中村 幸雄
		(72) 発明者	政瀧 浩和
			東京都千代田区大手町二丁目3番1号 日
			本電信電話株式会社内
		(72) 発明者	神 昭夫
			東京都千代田区大手町二丁目3番1号 日
			本電信電話株式会社内

最終頁に続く

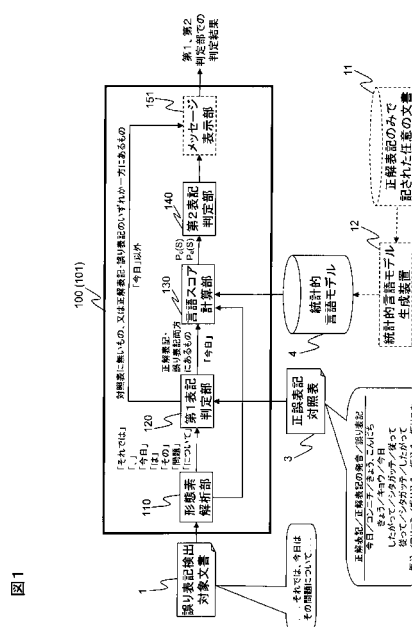
(54) 【発明の名称】 誤り表記検出装置、誤り表記生成装置、それらの方法、プログラムおよび記録媒体

(57) 【要約】

【課題】高い精度で表記のゆらぎや誤りを検出可能とする。

【解決手段】誤り表記検出対象文書 1 を形態素解析部 110 にて各単語に分解し、第 1 表記判定部 120 にて各単語について正誤表記対照表 3 と照合し、当該単語が正誤表記対照表 3 の正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には当該単語は正しい表記であると判定して結果を出力し、誤り表記のみに存在する場合には当該単語は誤った表記であると判定して結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、更に言語スコア計算部 130 にて、統計的言語モデル 4 を用いて各表記の言語スコアを計算し、第 2 表記判定部 140 にて各言語スコアの比較により正しい表記であるか誤った表記であるかを判定して結果を出力する。

【選択図】図1



【特許請求の範囲】**【請求項 1】**

誤り表記検出対象文書が入力され、単語単位に分割して単語列を出力する形態素解析部と、

上記単語列が入力され、当該単語列を構成する各単語について、正解表記とその正しい発音と当該発音の正解表記以外の表記（以下、「誤り表記」という）との組がリスト化された正誤表記対照表を参照して、当該単語の表記が正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には当該単語は正しい表記であるとの判定結果を出力し、誤り表記のみに存在する場合には当該単語は誤った表記であるとの判定結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、当該単語の表記と、当該単語の表記が誤り表記として含まれる組の正解表記（以下、「比較表記」という）とを出力する第 1 表記判定部と、

10

上記第 1 表記判定部にて上記単語の表記が上記正誤表記対照表の正解表記と誤り表記のいずれにも存在すると確認された場合に、上記単語列と当該単語の表記と上記比較表記とが入力され、当該単語列において当該単語の表記を、そのままにした場合の言語スコアと、当該比較表記に置き換えた場合の言語スコアを、正解表記のみで記された任意の文書から作成した統計的言語モデルを用いて、それぞれ計算して当該単語とともに出力する言語スコア計算部と、

上記言語スコア計算部で求めたそれぞれの言語スコアが入力され、上記単語の表記が正しい表記であるか誤った表記であるかの判定結果を出力する第 2 表記判定部と、
を備える誤り表記検出装置。

20

【請求項 2】

単語ごとに発音が付された誤り表記検出対象文書が入力され、各単語について、正解表記とその正しい発音と当該発音の正解表記以外の表記（以下、「誤り表記」という）との組がリスト化された正誤表記対照表を参照して、当該単語が正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には当該単語は正しい表記であるとの判定結果を出力し、誤り表記のみに存在する場合には誤った表記であるとの判定結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、当該単語とその発音を出力する第 1 表記判定部と、

上記第 1 表記判定部にて上記単語が上記正誤表記対照表の正解表記と誤り表記のいずれにも存在すると確認された場合に、上記単語とその発音が入力され、当該単語の発音について上記正誤表記対照表を参照して、その発音において当該単語が正解表記のみに存在する場合には正しい表記であるとの判定結果を出力し、誤り表記のみに存在する場合には誤った表記であるとの判定結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、当該単語と、当該単語が誤り表記として含まれる組の正解表記（以下、「比較表記」という）とを出力する発音判定部と、

30

上記発音判定部にて上記単語が上記正誤表記対照表の正解表記と誤り表記のいずれにも存在すると確認された場合に、上記誤り表記検出対象文書と当該単語と上記比較表記とが入力され、当該誤り表記検出対象文書において当該単語の表記を、そのままにした場合の言語スコアと、当該比較表記に置き換えた場合の言語スコアを、正解表記のみで記された任意の文書から作成した統計的言語モデルを用いて、それぞれ計算して当該単語とともに出力する言語スコア計算部と、

40

上記言語スコア計算部で求めたそれぞれの言語スコアが入力され、上記単語が正しい表記であるか誤った表記であるかの判定結果を出力する第 2 表記判定部と、
を備える誤り表記検出装置。

【請求項 3】

請求項 1 又は 2 のいずれかに記載の誤り表記検出装置において、

更に、上記判定結果が入力され、誤った表記であると判定された単語について、誤っている旨のメッセージを表示するか、誤り表記検出対象文書中にその他の単語と別の色で表示するか、又は正しい表記への訂正を促すメッセージを表示するメッセージ表示部を備え

50

ることを特徴とする誤り表記検出装置。

【請求項 4】

それぞれの単語に対して任意に定めた正解表記とその正しい発音との組がリスト化された正解表記・発音対照表の各正解表記について、それぞれの単語の表記とその読み方として想定される 1 以上の発音との組がリスト化された表記・発音対照辞書の同じ表記の組に係る発音を検索し、当該正解表記の発音以外の発音を当該正解表記の発音に追記して生成した第 1 中間リストを出力する表記同一発音検索部と、

上記第 1 中間リストが入力され、当該第 1 中間リストに含まれる上記正解表記の各発音をキーに、上記表記・発音対照辞書を検索して、当該各発音に対応する当該正解表記以外の表記を抽出し、抽出した表記を誤り表記候補として上記第 1 中間リストの当該正解表記に 10 関連付けて追加して生成した第 2 中間リストを出力する発音同一表記検索部と、

上記第 2 中間リストが入力され、当該第 2 中間リストの上記誤り表記候補から同音異義語を削除して、正解表記とその正しい発音と当該発音の正解表記以外の表記との組がリスト化された正誤表記対照表を生成して出力する同音異義語削除部と、
を備える誤り表記生成装置。

【請求項 5】

形態素解析部が、誤り表記検出対象文書を単語単位に分割して単語列を出力する形態素解析ステップと、

第 1 表記判定部が、上記単語列を構成する各単語について、正解表記とその正しい発音と当該発音の正解表記以外の表記（以下、「誤り表記」という）との組がリスト化された 20 正誤表記対照表を参照して、当該単語の表記が正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には当該単語は正しい表記であると判定して結果を出力し、誤り表記のみに存在する場合には当該単語は誤った表記であると判定して結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、当該単語の表記と、当該単語の表記が誤り表記として含まれる組の正解表記（以下、「比較表記」という）とを出力する第 1 表記判定ステップと、

言語スコア計算部が、上記第 1 表記判定ステップにて上記単語の表記が上記正誤表記対照表の正解表記と誤り表記のいずれにも存在すると確認された場合に、上記単語列において当該単語の表記を、そのままにした場合の言語スコアと、上記比較表記に置き換えた場合の言語スコアを、正解表記のみで記された任意の文書から作成した統計的言語モデルを 30 用いて、それぞれ計算して当該単語とともに出力する言語スコア計算ステップと、

第 2 表記判定部が、上記言語スコア計算ステップで求めたそれぞれの言語スコアから、上記単語の表記が正しい表記であるか誤った表記であるかを判定して結果を出力する第 2 表記判定ステップと、
を実行する誤り表記検出方法。

【請求項 6】

第 1 表記判定部が、単語ごとに発音が付された誤り表記検出対象文書の各単語について、正解表記とその正しい発音と当該発音の正解表記以外の表記（以下、「誤り表記」という）との組がリスト化された正誤表記対照表を参照して、当該単語が正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には当該単語は正しい表記で 40 であると判定して結果を出力し、誤り表記のみに存在する場合には誤った表記であると判定して結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、当該単語とその発音を出力する第 1 表記判定ステップと、

発音判定部が、上記第 1 表記判定ステップにて上記単語が上記正誤表記対照表の正解表記と誤り表記のいずれにも存在すると確認された場合に、上記単語の発音について上記正誤表記対照表を参照して、その発音において当該単語が正解表記のみに存在する場合には正しい表記であると判定して結果を出力し、誤り表記のみに存在する場合には誤った表記であると判定して結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、当該単語と、当該単語が誤り表記として含まれる組の正解表記（以下、「比較表記」という）とを出力する発音判定ステップと、

10

20

30

40

50

言語スコア計算部が、上記発音判定ステップにて上記単語が上記正誤表記対照表の正解表記と誤り表記のいずれにも存在すると確認された場合に、上記誤り表記検出対象文書において当該単語の表記を、そのままにした場合の言語スコアと、上記比較表記に置き換えた場合の言語スコアを、正解表記のみで記された任意の文書から作成した統計的言語モデルを用いて、それぞれ計算して当該単語とともに出力する言語スコア計算ステップと、

第2表記判定部が、上記言語スコア計算ステップで求めたそれぞれの言語スコアから、上記単語の表記が正しい表記であるか誤った表記であるかを判定して結果を出力する第2表記判定ステップと、
を実行する誤り表記検出方法。

【請求項7】

10

請求項5又は6のいずれかに記載の誤り表記検出方法において、

更に、メッセージ表示部が、上記判定の結果、誤った表記であると判定された単語について、誤っている旨のメッセージを表示するか、誤り表記検出対象文書中にその他の単語と別の色で表示するか、又は正しい表記への訂正を促すメッセージを表示するメッセージ表示ステップを実行することを特徴とする誤り表記検出方法。

【請求項8】

表記同一発音検索部が、それぞれの単語に対して任意に定めた正解表記とその正しい発音との組がリスト化された正解表記・発音対照表の各正解表記について、それぞれの単語の表記とその読み方として想定される1以上の発音との組がリスト化された表記・発音対照辞書の同じ表記の組に係る発音を検索し、当該正解表記の正しい発音以外の発音を当該正解表記の発音として追記して生成した第1中間リストを出力する表記同一発音検索ステップと、

20

発音同一表記検索部が、上記第1中間リストに含まれる上記正解表記の各発音をキーに、上記表記・発音対照辞書を検索して、当該各発音に対応する当該正解表記以外の表記を抽出し、抽出した表記を誤り表記候補として上記第1中間リストの当該正解表記に関連付けて追加して生成した第2中間リストを出力する発音同一表記検索ステップと、

同音異義語削除部が、上記第2中間リストの上記誤り表記候補から同音異義語を削除して、正解表記とその正しい発音と当該発音の正解表記以外の表記との組がリスト化された正誤表記対照表を生成して出力する同音異義語削除ステップと、
を実行する誤り表記生成方法。

30

【請求項9】

請求項1～4のいずれかに記載した装置としてコンピュータを機能させるためのプログラム。

【請求項10】

請求項9に記載したプログラムを記録したコンピュータが読み取り可能な記録媒体。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、文書中の表記を推敲する技術に関し、具体的には表記のゆらぎ、誤りを正確に統一、修正するための誤り表記検出装置、誤り表記生成装置、それらの方法、プログラムおよび記録媒体に関する。

40

【背景技術】

【0002】

法律に関わる文書や議会の議事録など、表記を正確に統一する必要がある文書をパーソナルコンピュータやワードプロセッサなどを用いて作成する際、表記のゆらぎや誤りを自動検出する方法として、従来、以下のような手法が実用化されている。

(1)任意に定めた統一基準による正解表記と誤り表記との組情報を予め持ち、文書から誤り表記が検出された場合に、その組情報に基づき訂正候補を提示する手法。

(2)ルールを作成して外来語や送り仮名などのゆらぎを訂正する手法。例えば、「ター」を「タ」に(ex.「インターフェース」→「インタフェース」)、「込み」を「込」に(e

50

x. 「振込み」→「振込」)、訂正するというルールを作成する。

【特許文献1】特開平5-233620号公報

【発明の開示】

【発明が解決しようとする課題】

【0003】

上記の従来手法には次のような問題がある。

(1)表記のみで正解、誤りを判断しているため、同じ表記であるにもかかわらず、ある読み方としては誤りであり、ある読み方としては正解であるような場合に対応できない。例えば、「今日」という表記がある場合、「キョウ」と読みたい場合には「きょう」が正解表記(つまり「今日」は誤り表記)であり、「コンニチ」と読みたい場合には「今日」が正解表記であると統一基準として定めた時、文書に「今日」という表記が現われた時に「きょう」に訂正すべきか「今日」のままよいかの切り分けができない。

【0004】

また、正解表記に対する誤り表記は必ずしも一つではないため、誤り候補を予め準備するための準備に手間がかかる。

(2)外来語や送り仮名以外にも表記を統一する必要がある、また統一の規則は単語ごとに異なるため、例えば、必ずしも全ての「ター」を「タ」に訂正すべきであるとは限らず、ルールだけでは限界がある。

【0005】

本発明の目的は、同じ表記であるにもかかわらず、ある読み方としては誤りであり、ある読み方としては正解であるような場合にも、高い精度で表記のゆらぎや誤りを自動検出でき、また、誤りの候補やルールを作成する必要が無い、誤り表記検出装置、誤り表記生成装置、それらの方法、プログラムおよび記録媒体を提供することにある。

【課題を解決するための手段】

【0006】

本発明の誤り表記検出装置は、形態素解析部と第1表記判定部と言語スコア計算部と第2表記判定部とを備える。形態素解析部は、誤り表記検出対象文書が入力され、単語単位に分割して単語列を出力する。第1表記判定部は、当該単語列が入力され、当該単語列を構成する各単語について、正解表記とその正しい発音と当該発音の正解表記以外の表記(以下、「誤り表記」という)との組がリスト化された正誤表記対照表を参照して、当該単語が正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には当該単語は正しい表記であるとの判定結果を出力し、誤り表記のみに存在する場合には当該単語は誤った表記であるとの判定結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、当該単語と当該単語が誤り表記として含まれる組の正解表記(以下、「比較表記」という)とを出力する。言語スコア計算部は、第1表記判定部にて当該単語が正誤表記対照表の正解表記と誤り表記のいずれにも存在すると確認された場合に、当該単語列と当該単語と当該比較表記とが入力され、当該単語列において当該単語の表記を、そのままにした場合の言語スコアと、当該比較表記に置き換えた場合の言語スコアを、正解表記のみで記された任意の文書から作成した統計的言語モデルを用いて、それぞれ計算して当該単語とともに出力する。第2表記判定部は、言語スコア計算部で求めたそれぞれの言語スコアが入力され、当該単語が正しい表記であるか誤った表記であるかの判定結果を出力する。

【0007】

また、本発明の誤り表記生成装置は、表記同一発音検索部と発音同一表記検索部と同音異義語削除部とを備える。表記同一発音検索部は、それぞれの単語に対して任意に定めた正解表記とその正しい発音との組がリスト化された正解表記・発音対照表の各正解表記について、それぞれの単語の表記とその読み方として想定される1以上の発音との組がリスト化された表記・発音対照辞書の同じ表記の組に係る発音を検索し、当該正解表記の正しい発音以外の発音を当該正解表記の発音に追記して生成した第1中間リストを出力する。発音同一表記検索部は、第1中間リストが入力され、第1中間リストに含まれる当該正解

10

20

30

40

50

表記の各発音をキーに、表記・発音対照辞書を検索して、当該各発音に対応する当該正解表記以外の表記を抽出し、抽出した表記を誤り表記候補として第１中間リストの当該正解表記に関連付けて追加して生成した第２中間リストを出力する。同音異義語削除部は、第２中間リストが入力され、第２中間リストの当該誤り表記候補から同音異義語を削除して、正解表記とその正しい発音と当該発音の正解表記以外の表記との組がリスト化された正誤表記対照表を生成して出力する。

【発明の効果】

【０００８】

本発明の誤り表記検出装置と誤り表記生成装置によれば、表記に加えて発音も検出に用いるため、同じ表記であるにもかかわらず、ある読み方としては誤りであり、ある読み方としては正解である場合にも、高い精度で表記のゆらぎや誤りを自動検出でき、また、誤りの候補やルールを作成する必要が無くなり、文書編集作業の負担を軽減することができる。

10

【発明を実施するための最良の形態】

【０００９】

〔第１実施形態〕

図１に本発明の誤り表記検出装置１００の機能構成例を、図２にその処理フロー例を示す。誤り表記検出装置１００は、形態素解析部１１０と第１表記判定部１２０と言語スコア計算部１３０と第２表記判定部１４０とから構成され、入力された誤り表記検出対象文書１の各単語それぞれについて正解表記であるか誤り表記であるかを判定し結果を出力する。

20

【００１０】

形態素解析部１１０は、誤り表記を検出する対象文書である、 n 個の単語からなる誤り表記検出対象文書１が入力され、文書を単語単位に分割して単語列 w_1 、 w_2 、 $\dots$ 、 w_n を出力する（Ｓ１）。例えば、図２に示すように「それでは、今日はその問題について」という文書が入力された時、「それでは」「、」「今日」「は」「その」「問題」「について」というように分割した単語列を出力する。

【００１１】

第１表記判定部１２０は、形態素解析部１１０で分割された単語列 w_1 、 w_2 、 $\dots$ 、 w_n が入力され、それぞれの単語 w_k （ $1 \leq k \leq n$ ）について、正解表記とその正しい発音と当該発音の正解表記以外の表記（誤り表記）との組がリスト化された正誤表記対照表３を参照して、単語 w_k が正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には正しい表記であると判定し、誤り表記のみに存在する場合には誤った表記であると判定して、それぞれの単語ごとに判定結果を出力する。また、単語 w_k が正解表記と誤り表記のいずれにも存在する場合には、単語 w_k と、単語 w_k が誤り表記として含まれる組の正解表記（比較表記） w_k' とを出力する（Ｓ２）。例えば、「それでは」「、」「今日」「は」「その」「問題」「について」という単語列が入力された場合において、正誤表記対照表３が図２に示すような内容である時、正解表記と誤り表記のいずれにも存在するのは「今日」のみであり、その他の単語はいずれにも存在しないため、「今日」と、「今日」が誤り表記として含まれる組の正解表記（比較表記）である「きょう」とを言語スコア計算部１３０に向けて出力し、それ以外の単語については正しい表記であるとの判定結果を出力する。

30

40

【００１２】

なお、ここで参照する正誤表記対照表３は、手作業で作成することも可能であるが、正解表記に対する誤り表記は必ずしも一つではないため、誤り候補を予め準備するための準備に手間がかかる。そのため、正誤表記対照表３を後述する誤り表記生成装置３００により生成することで準備を省力化を図ることで、本発明による誤り表記検出処理をより円滑に行うことが可能となる。

【００１３】

言語スコア計算部１３０は、第１表記判定部１２０にて単語 w_k が上記正誤表記対照表

50

の正解表記と誤り表記のいずれにも存在すると確認された場合に、形態素解析部 110 で分割された単語列 $w_1, w_2, \dots, w_n$ と、単語 w_k と上記比較表記 w_k' とが入力され、当該単語列において単語 w_k をそのままにした場合の言語スコア $P_c(S)$ と、比較表記 w_k' に置き換えた場合の言語スコア $P_e(S)$ をそれぞれ、統計的言語モデル 4 を用いて計算して単語 w_k とともに出力する (S3)。

【0014】

$$P_c(S) = P(w_1, \dots, w_{k-2}, w_{k-1}, w_k, w_{k+1}, w_{k+2}, \dots, w_n)$$

$$P_e(S) = P(w_1, \dots, w_{k-2}, w_{k-1}, w_k', w_{k+1}, w_{k+2}, \dots, w_n)$$

統計的言語モデルとは、与えられた文書に対して、その文書を構成する単語列の生成確率を求めるためのモデルである。統計的言語モデルには連続音声認識などで盛んに使用されている N-gram を始めとして、任意の統計的言語モデルを使用できるが、ここでは Trigram を使用した場合を例にとって言語スコアの計算方法を説明する。

Trigram は、単語列中のある単語 w_t の出現確率を直前の 2 単語から予測するモデルで $P(w_t | w_{t-2}, w_{t-1})$ として表現される。従って、文書全体の言語スコアは次式により求めることができる。

【0015】

【数 1】

$$P(S) = \prod_{t=1}^n P(w_t | w_{t-2}, w_{t-1}) \quad (1)$$

また、確率の掛け算により値が小さくなるため、各確率の対数の和として次式のように求めてもよい。

【0016】

【数 2】

$$P'(S) = \prod_{t=1}^n \log P(w_t | w_{t-2}, w_{t-1}) \quad (2)$$

もっとも、本発明における言語スコア計算部 130 では、判定する単語 w_k に関わる箇所の言語スコアのみがわかれば判定可能なため、次式のように積算を省略しても構わない。

【0017】

【数 3】

$$P''(S) = \prod_{t=k}^{k+2} \log P(w_t | w_{t-2}, w_{t-1}) \quad (3)$$

統計的言語モデル 4 ($P(w_t | w_{t-2}, w_{t-1})$) は、言語スコア計算部 230 での言語スコアの計算に先立ち、正解表記のみで記された任意の文書 11 を用いて、公知の統計的言語モデル生成方法 (例えば、参考文献 1、2 参照) を実行する統計的言語モデル生成装置 12 にて事前に生成しておく。

【0018】

【参考文献 1】中川聖一著、電子情報通信学会編、「確率モデルによる音声認識」、第 1 版、株式会社コロナ社、1988 年 7 月、p.113-121

【参考文献 2】鹿野清宏、外 4 名編著、情報処理学会編、「音声認識システム」、第 1 版、株式会社オーム社、2001 年 5 月、p.53-68

統計的言語モデル 4 の生成に用いる正解表記のみで記された任意の文書 11 は、本発明での主な誤り表記検出対象文書である法律に係わる文書や議会の議事録が標準的な用字用語に則るものであることから、そのような既存の文書や議会の議事録を適宜選定してそれをそのまま利用することができる。また、参考文献 3 に記載された文部省用字用語例や参考文献 4 の用字用語辞典に記載された標準的な用字用語を正解表記として利用し、これに基づいて任意の文書 11 を容易に構成することもできる。特に、議会の議事録が対象であ

10

20

30

40

50

る場合は、参考文献 5、6 を正解表記として利用し、これに基づいて任意の文書 11 を構成すればよい。

【0019】

〔参考文献 3〕文化庁編、「公用文の書き表し方の基準（資料集）」、増補 2 版、第一法規株式会社、2001 年 5 月

〔参考文献 4〕NHK 放送文化研究所編、「NHK 新用字用語辞典」、第 3 版、株式会社日本放送出版協会、2004 年 3 月

〔参考文献 5〕衆議院事務局記録部、参議院事務局記録部編、「国会会議録用字例」、衆議院事務局記録部、参議院事務局記録部、1975 年

〔参考文献 6〕日本速記協会編、「改訂 標準用字用例辞典」、社団法人日本速記協会、2007 年

10

なお、正解表記のみで記された任意の文書 11 のデータ量は多ければ多いに越したことはないが、一例として、のべ約 650 時間、単語（形態素）にして約 700 万語規模を目安とすることが考えられる（参考文献 7 参照）。

【0020】

〔参考文献 7〕古井貞熙、「一里塚としての「日本語話し言葉コーパス」」、[online]、2006 年 3 月、日本音響学会講演論文集、p.1192、[2008 年 9 月 3 日検索]、インターネット <URL: http://www.furui.cs.titech.ac.jp/publication/2006/0232_3-1-4.pdf>

第 2 表記判定部 140 は、単語 w_k と、当該単語 w_k について求めた言語スコア $P_c(S)$ 、 $P_e(S)$ が入力され、単語 w_k の表記が正しい表記であるか誤った表記であるかを判定して、判定結果を出力する（S4）。判定方法は、例えば単純に、 $P_e(S) > P_c(S)$ である場合に単語 w_k の表記は誤っていると判定してもよいし、正誤判定の性能を調整するために、 $P_e(S) > \alpha \cdot P_c(S)$ （ α は正の定数）である場合に単語 w_k の表記は誤っていると判定してもよい。後者の場合、 α が小さいほど単語を誤りとして検出しやすくなる。なお、単語 w_k が誤り表記として含まれる組が複数ある場合は、それぞれの組の比較表記 w_k' について言語スコア $P_e(S)$ を算出し、算出した全ての $P_e(S)$ のうち最もスコアが高いものを選択して上記のように $P_c(S)$ との比較判定を行う。

20

以上のように誤り表記検出装置を構成することで、誤り検出を表記に加えて発音も用いて行うため、同じ表記であるにもかかわらず、ある読み方としては誤りであり、ある読み方としては正解であるような場合にも、高い精度で表記のゆらぎや誤りを自動検出することができる。

30

【0021】

〔第 2 実施形態〕

図 3 に本発明の誤り表記検出装置 200 の機能構成例を、図 4 にその処理フロー例を示す。誤り表記検出装置 200 は、第 1 表記判定部 220 と発音判定部 260 と言語スコア計算部 230 と第 2 表記判定部 140 とから構成され、入力された誤り表記検出対象文書 2 の各単語それぞれについて正解表記であるか誤り表記であるかを判定し結果を出力する。

【0022】

第 2 実施形態は、音声認識の結果などにより、誤り表記検出対象文書の各単語に対して実際の発音が付与されている場合に好適な構成である。第 1 実施形態の機能構成との主な相違は、誤り表記検出対象文書が予め単語単位に分割されているため形態素解析部 110 が無い一方、言語スコアの計算を行うか否かを、誤り表記検出対象文書の各単語に付与された実際の発音に応じて決定するための発音判定部 260 が加わっている点が挙げられる。

40

【0023】

第 1 表記判定部 220 は、誤り表記を検出する対象文書である、 n 個の単語 w_1 、 w_2 、 $\dots$ 、 w_n からなる単語ごとに発音が付与された誤り表記検出対象文書 2 が入力され、それぞれの単語 w_k （ $1 \leq k \leq n$ ）について、正解表記とその正しい発音と当該発音の正解表記以外の表記（誤り表記）との組がリスト化された正誤表記対照表 3 を参照して、

50

単語 w_k が正解表記のみに存在するか又は正解表記と誤り表記のいずれにも存在しない場合には正しい表記であると判定し、誤り表記のみに存在する場合には誤った表記であると判定して、それぞれの単語ごとに判定結果を出力する。また、単語 w_k が正解表記と誤り表記のいずれにも存在する場合には、単語 w_k とその発音を発音判定部 260 に向けて出力する (S5)。その他の説明は第1実施形態の第1表記判定部 120 の所で記したとおりである。

【0024】

発音判定部 260 は、第1表記判定部 220 にて単語 w_k が正誤表記対照表 3 の正解表記と誤り表記のいずれにも存在すると確認された場合に、単語 w_k とその発音とが入力され、単語 w_k の発音について正誤表記対照表 3 を参照して、その発音において単語 w_k が正解表記のみに存在する場合には正しい表記であるとの判定結果を出力し、誤り表記のみに存在する場合には誤った表記であるとの判定結果を出力し、正解表記と誤り表記のいずれにも存在する場合には、単語 w_k と、単語 w_k が誤り表記として含まれる組の正解表記 (比較表記) w_k' とを出力する (S6)。例えば、図 4 に示すように単語 w_k が「今日」でその発音が「キョウ」である場合、正誤表記対照表 3 において「キョウ」の発音に対する表記「今日」は誤り表記のみに存在するため、誤った表記であると判定結果を出力する。

10

【0025】

言語スコア計算部 230 は、発音判定部 260 で単語 w_k が正誤表記対照表 3 の正解表記と誤り表記のいずれにも存在すると確認された場合に、誤り表記検出対象文書 2 と単語 w_k と比較表記 w_k' とが入力され、誤り表記検出対象文書 2 において単語 w_k の表記をそのままにした場合の言語スコア $P_o(S)$ と、比較表記 w_k' に置き換えた場合の言語スコア $P_e(S)$ をそれぞれ、統計的言語モデル 4 を用いてそれぞれ計算して単語 w_k とともに出力する (S7)。なお、言語スコアの計算方法や統計的言語モデル 4 の生成方法については、第1実施形態の言語スコア計算部 130 の所で説明したとおりである。

20

【0026】

第2表記判定部 140 は、第1実施形態と同様のものであり、単語 w_k と、単語 w_k について求めた言語スコア $P_o(S)$ 、 $P_e(S)$ とが入力され、単語 w_k の表記が正しい表記であるか誤った表記であるかを判定して、判定結果を出力する (S4)。

以上のように誤り表記検出装置を構成することで、第1実施形態と同様、誤り検出を表記に加えて発音も用いて行うため、同じ表記であるにもかかわらず、ある読み方としては誤りであり、ある読み方としては正解であるような場合にも、高い精度で表記のゆらぎや誤りを自動検出することができる。

30

【0027】

〔第3実施形態〕

第3実施形態の誤り表記検出装置 101 及び 201 は、図 1 及び図 3 の点線で示すように、第1実施形態の誤り表記検出装置 100 及び第2実施形態の誤り表記検出装置 200 の判定結果出力端に、それぞれメッセージ表示部 151 を接続した構成である。メッセージ表示部 151 は、第1表記判定部 120 (220)、第2表記判定部 140 及び発音判定部 260 での判定結果が入力され、誤った表記であると判定された単語 w_k について、誤っている旨のメッセージを表示するか、誤り表記検出対象文書全体を表示する中でその他の単語と別の色で表示するか、又は正しい表記への訂正を促すメッセージを表示する。

40

このようにメッセージを表示することで、表記の誤りを直接視認することができる。

【0028】

〔第4実施形態〕

上記各実施形態は誤り表記検出装置に係るものであったが、第4実施形態は上記各実施形態における誤り表記検出処理の中で用いる正誤表記対照表 3 を生成する誤り表記生成装置 300 に係るものである。図 5 に本発明の誤り表記生成装置 300 の機能構成例を、図 6 にその処理フロー例を示す。また、誤り表記検出装置 100 (200) と誤り表記生成装置 300 とを組み合わせた全体構成イメージを図 7 に示す。

50

【 0 0 2 9 】

誤り表記生成装置 3 0 0 は、表記同一発音検索部 3 1 0 と発音同一表記検索部 3 2 0 と同音異義語削除部 3 3 0 とから構成される。

【 0 0 3 0 】

誤り表記生成装置 3 0 0 における処理に先立ち、それぞれの単語に対して任意のルールに基づき定めた正解表記とその正しい発音との組がリスト化された正解表記・発音対照表 2 1 と、それぞれの単語の表記とその読み方として想定される 1 以上の発音との組がリスト化された表記・発音対照辞書 2 2 とをあらかじめ用意しておく。

【 0 0 3 1 】

なお、正解表記・発音対照表 2 1 は、本発明での主な誤り表記検出対象文書である法律に係わる文書や議会の議事録が、標準的な用字用語に則るものであることから、例えば、参考文献 3 に記載された文部省用字用語例や参考文献 4 の用字用語辞典に記載された標準的な用字用語をベースに効率的に作成することができる。また特に、議会の議事録が対象である場合は、参考文献 5、6 を正解表記・発音対照表 2 1 としてそのまま利用することができる。

10

【 0 0 3 2 】

また、表記・発音対照辞書 2 2 は、特別なものである必要はなく、一般的な電子辞書、形態素解析・かな漢字用の辞書など、単語の表記と発音とが対になったものであれば構わない。

【 0 0 3 3 】

20

表記同一発音検索部 3 1 0 は、正解表記・発音対照表 2 1 の各正解表記について、表記・発音対照辞書 2 2 の同じ表記の組に係る発音を検索し、当該正解表記の正しい発音以外の発音を当該正解表記の発音として追記して第 1 中間リスト 2 3 を生成・出力する（S 1 1）。例えば、図 5 に示すように正解表記・発音対照表 2 1 の正解表記が「今日」でその正しい発音が「コンニチ」である場合には、表記・発音対照辞書 2 2 の表記「今日」に対する発音が「キョウ」「コンニチ」であることから、差分である「キョウ」を追記し、正解表記「今日」に対応する発音を「キョウ」「コンニチ」として第 1 中間リスト 2 3 を生成する。

【 0 0 3 4 】

発音同一表記検索部 3 2 0 は、第 1 中間リスト 2 3 が入力され、第 1 中間リスト 2 3 に含まれる正解表記の各発音をキーに、表記・発音対照辞書 2 2 を検索して、当該各発音に対応する当該正解表記以外の表記を抽出し、抽出した表記を誤り表記候補として第 1 中間リスト 2 3 の当該正解表記に関連付けて追加して、第 2 中間リスト 2 4 を生成・出力する（S 1 2）。例えば、正解表記「今日」に対応する発音「キョウ」「コンニチ」のそれぞれについて、表記・発音対照辞書 2 2 を検索すると、「キョウ」については「今日」「きょう」が、「コンニチ」については「今日」「こんにち」が対応する表記であるため、正解表記である「今日」を除いた「きょう」「こんにち」を、正解表記「今日」に対する誤り表記候補として第 1 中間リスト 2 3 の当該正解表記に関連付けて追加し、第 2 中間リスト 2 4 を生成・出力する。

30

【 0 0 3 5 】

40

同音異義語削除部 3 3 0 は、第 2 中間リスト 2 4 が入力され、第 2 中間リスト 2 4 の誤り表記候補から同音異義語を削除して、正解表記とその正しい発音と当該発音の正解表記以外の表記（誤り表記）との組がリスト化された正誤表記対照表 3 を生成して出力する（S 1 3）。例えば、正解表記が「振込」の場合、その発音が「フリコミ」であり、誤り表記候補には「振り込み」「振込み」「振りこみ」「降りこみ」「降込み」があるが、「降りこみ」「降込み」は同音異義語であるため削除し、残った「振り込み」「振込み」「振りこみ」が、正解表記「振込」に対する誤り表記となる。

【 0 0 3 6 】

以上のように誤り表記生成装置を構成することで、実質的に正解表記を準備するだけで誤り表記を生成することができるため、誤りの候補やルールを作成する必要がなくなり、

50

文書編集作業の負担を軽減することができる。

【 0 0 3 7 】

上記の各実施形態の誤り表記検出装置、誤り表記生成装置の構成をコンピュータによって実現する場合、各装置が有すべき機能の処理内容はプログラムによって記述される。そして、このプログラムをコンピュータで実行することにより、上記処理機能がコンピュータ上で実現される。

【 0 0 3 8 】

この処理内容を記述したプログラムは、コンピュータで読み取り可能な記録媒体に記録しておくことができる。コンピュータで読み取り可能な記録媒体としては、例えば、磁気記録装置、光ディスク、光磁気記録媒体、半導体メモリ等のようなものでもよいが、具体的には、例えば、磁気記録装置として、ハードディスク装置、フレキシブルディスク、磁気テープ等を、光ディスクとして、D V D (Digital Versatile Disc)、D V D - R A M (Random Access Memory)、C D - R O M (Compact Disc Read Only Memory)、C D - R (Recordable) / R W (ReWritable)等を、光磁気記録媒体として、M O (Magneto-Optical disc)等を、半導体メモリとしてE E P - R O M (Electrically Erasable and Programmable-Read Only Memory)等を用いることができる。

10

【 0 0 3 9 】

また、このプログラムの流通は、例えば、そのプログラムを記録したD V D、C D - R O M等の可搬型記録媒体を販売、譲渡、貸与等することによって行う。さらに、このプログラムをサーバコンピュータの記憶装置に格納しておき、ネットワークを介して、サーバコンピュータから他のコンピュータにそのプログラムを転送することにより、このプログラムを流通させる構成としてもよい。

20

【 0 0 4 0 】

また、上述した実施形態とは別の実行形態として、コンピュータが可搬型記録媒体から直接このプログラムを読み取り、そのプログラムに従った処理を実行することとしてもよく、さらに、このコンピュータにサーバコンピュータからプログラムが転送されるたびに、逐次、受け取ったプログラムに従った処理を実行することとしてもよい。また、サーバコンピュータから、このコンピュータへのプログラムの転送は行わず、その実行指示と結果取得のみによって処理機能を実現する、いわゆるA S P (Application Service Provider)型のサービスによって、上述の処理を実行する構成としてもよい。なお、本形態におけるプログラムには、電子計算機による処理の用に供する情報であってプログラムに準ずるもの(コンピュータに対する直接の指令ではないがコンピュータの処理を規定する性質を有するデータ等)を含むものとする。

30

【 0 0 4 1 】

また、この形態では、コンピュータ上で所定のプログラムを実行させることにより、本装置を構成することとしたが、これらの処理内容の少なくとも一部をハードウェア的に実現することとしてもよい。

【 0 0 4 2 】

また、上述の各種の処理は、記載に従って時系列に実行されるのみならず、処理を実行する装置の処理能力あるいは必要に応じて並列的にあるいは個別に実行されてもよい。その他、本発明の趣旨を逸脱しない範囲で適宜変更が可能であることはいうまでもない。

40

【産業上の利用可能性】

【 0 0 4 3 】

本発明は、文書を推敲が必要な用途において、表記のゆらぎ、誤りをより正確に統一、修正したい場合に有用である。

【図面の簡単な説明】

【 0 0 4 4 】

【図 1】第 1、3 実施形態の誤り表記検出装置 1 0 0、1 0 1 の機能構成例を示す図

【図 2】第 1、3 実施形態の誤り表記検出装置 1 0 0、1 0 1 処理フロー例を示す図

【図 3】第 2、3 実施形態の誤り表記検出装置 2 0 0、2 0 1 の機能構成例を示す図

50

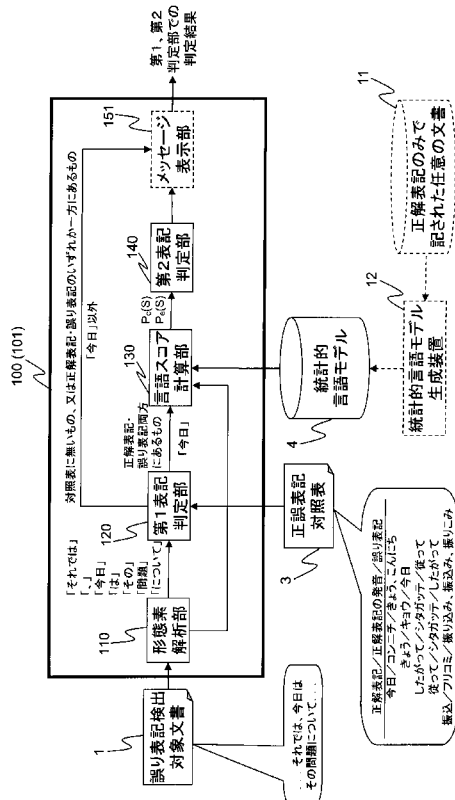
【図4】第2、3実施形態の誤り表記検出装置200、201処理フロー例を示す図

【図5】第4実施形態の誤り表記生成装置300の機能構成例を示す図

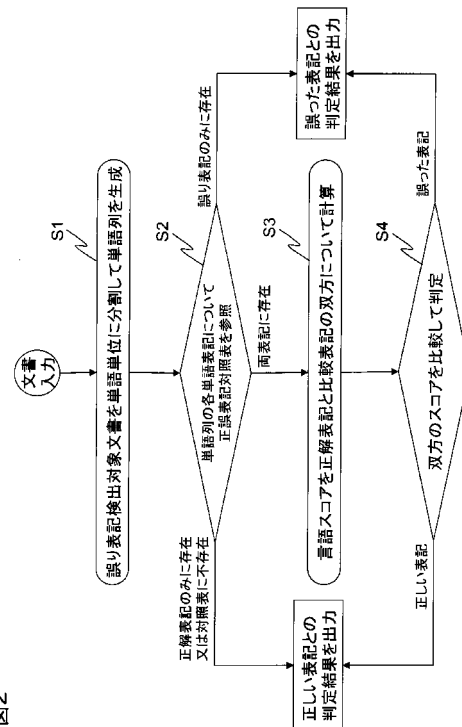
【図6】第4実施形態の誤り表記生成装置300の処理フロー例を示す図

【図7】誤り表記検出装置100、200と誤り表記生成装置300との組み合わせイメージを示す図

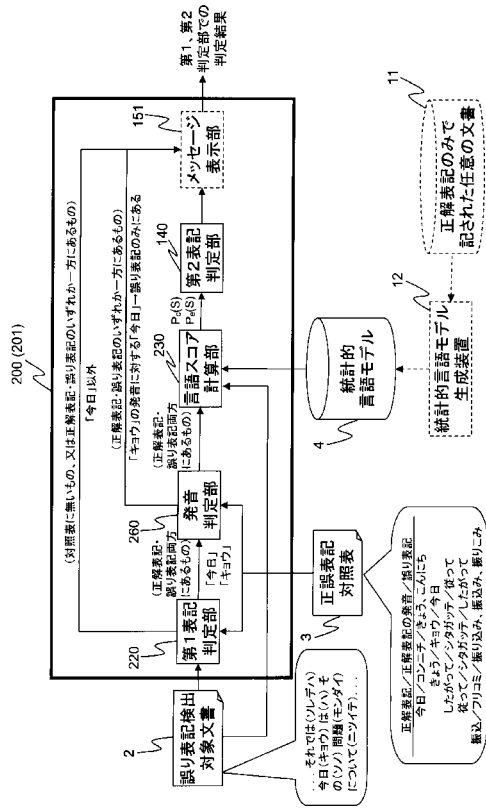
【図1】



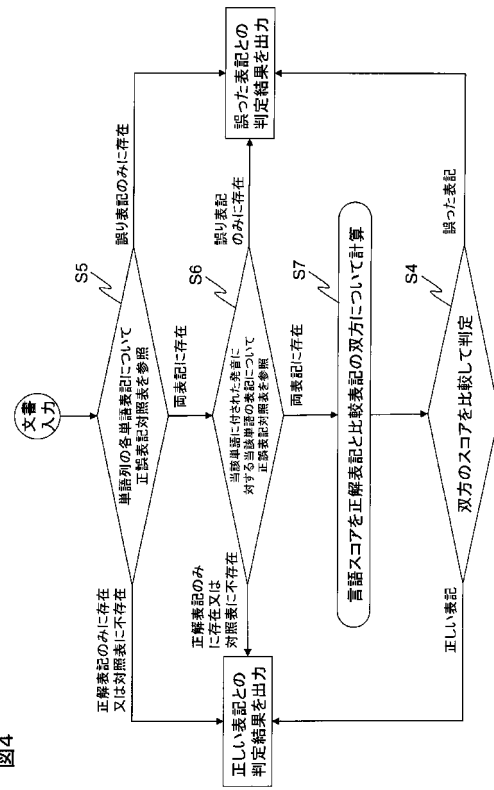
【図2】



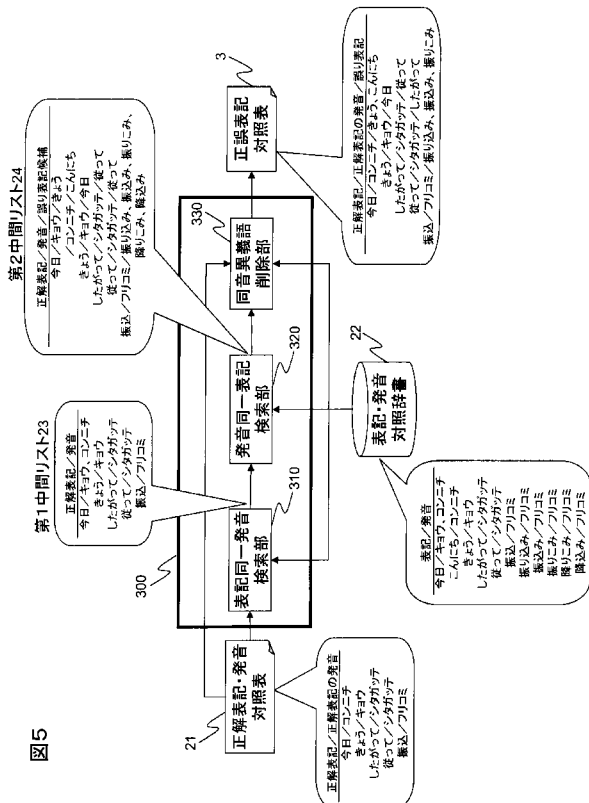
【図 3】



【図 4】



【図 5】



【図 6】

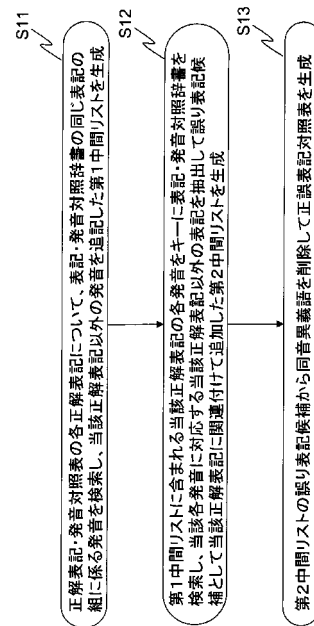
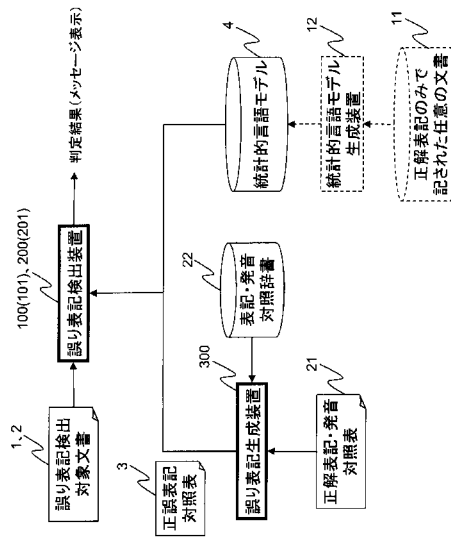


図7



フロントページの続き

(72)発明者 高橋 敏

東京都千代田区大手町二丁目 3 番 1 号 日本電信電話株式会社内

F ターム(参考) 5B109 KA05 QA01 QA14