

(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2015-130102

(P2015-130102A)

(43) 公開日 平成27年7月16日(2015.7.16)

(51) Int.Cl.	F I	テーマコード (参考)
G 0 6 F 17/21 (2006.01)	G 0 6 F 17/21 5 5 0 A	5 B 0 9 1
G 0 6 F 17/27 (2006.01)	G 0 6 F 17/21 5 5 0 J	5 B 1 0 9
	G 0 6 F 17/27 Z	

審査請求 未請求 請求項の数 10 O L (全 23 頁)

(21) 出願番号	特願2014-1641 (P2014-1641)	(71) 出願人	000004237
(22) 出願日	平成26年1月8日 (2014.1.8)		日本電気株式会社
			東京都港区芝五丁目7番1号
		(74) 代理人	100109313
			弁理士 机 昌彦
		(74) 代理人	100124154
			弁理士 下坂 直樹
		(72) 発明者	平尾 英司
			東京都港区芝五丁目7番1号
			日本電気株式会社内
		Fターム(参考)	5B091 CA00
			5B109 KA05 QA01 QA14

(54) 【発明の名称】 文書分析装置、文書分析システム、文書分析方法およびプログラム

(57) 【要約】

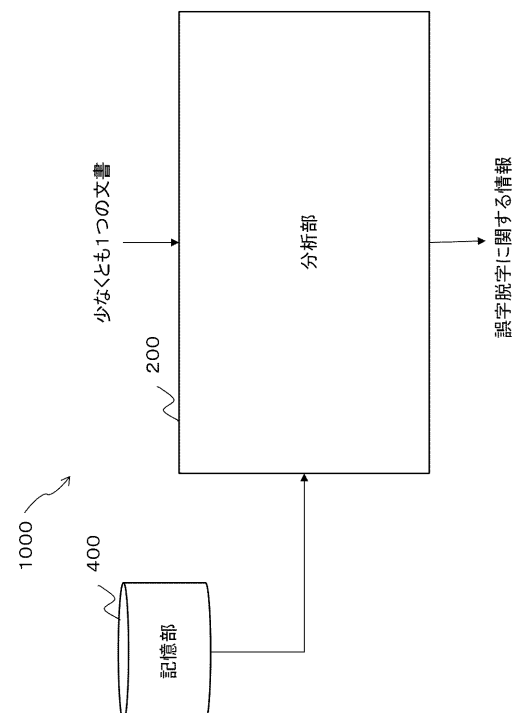
【課題】

文書の内容に依存する誤字脱字を抽出することができない。

【解決手段】

本発明の文書分析装置は、誤字脱字を含む単語のパターン情報を記憶する記憶部と、少なくとも1つの文書から前記文書が有する単語に関する情報を抽出し、当該単語に関する情報と前記パターン情報とに基づいて、誤字脱字に関する情報を抽出する分析部と、を含む。

【選択図】 図11



【特許請求の範囲】

【請求項 1】

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶する記憶部と、

少なくとも 1 つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する分析部と、
を含む文書分析装置。

【請求項 2】

前記誤字脱字に関する情報は、少なくとも前記文書が有する正しい単語と、前記正しい単語に対応する誤字脱字を含む単語と、前記誤字脱字を含む単語に対する誤字脱字の可能性を示す前記誤字脱字指標との対応関係を含み、

前記分析部は、

前記誤字脱字に関する情報から抽出する前記対応関係の数を算出し、算出した前記対応関係の数を前記誤字脱字に関する情報に含め、前記誤字脱字指標の値の大きい順に、前記対応関係の数だけ前記対応関係を抽出する請求項 1 に記載の文書分析装置。

【請求項 3】

前記分析部は、文書解析部と、誤字脱字指標生成部と、ノイズ情報検出部と、誤字脱字抽出条件変更部と、誤字脱字抽出部とをさらに有し、前記記憶部はノイズデータベースをさらに有し、

前記文書解析部は、前記文書が有する少なくとも 1 つの文章に形態素解析を適用することで前記単語情報を抽出し、

前記誤字脱字指標生成部は、前記単語情報から前記誤字脱字指標を算出し、

前記ノイズデータベースは、前記誤字脱字指標で前記誤字脱字を含む単語を検出しようとした場合に誤検出となるパターンを抽出可能にする情報を蓄積し、任意の単語の組み合わせが誤検出パターンに該当するかどうかに関する問い合わせに対し、問い合わせ対象の単語の組み合わせが誤検出パターンと一致するか否かを示す情報を応答し、

前記ノイズ情報検出部は、前記誤字脱字指標に基づき、複数の前記単語の組み合わせに対して、前記誤字脱字指標で誤検出となるパターンの有無を前記ノイズデータベースに問合せ、ノイズとなる前記単語の組み合わせをノイズ情報として検出し、

前記誤字脱字抽出条件変更部は、前記ノイズ情報を利用して所定の誤字・脱字抽出条件変更ルールに基づき、前記誤字脱字を含む単語の抽出条件を変更し、

前記誤字脱字抽出部は、前記文書内の全ての前記単語の組み合わせから、前記ノイズ情報に該当する前記単語の組み合わせを除外し、前記抽出条件に基づき、前記誤字脱字を含む単語を前記誤字脱字に関する情報として抽出する請求項 1 又は 2 に記載の文書分析装置。

【請求項 4】

前記誤字脱字指標は、事前に前記文書で使用可能な単語の一覧を登録した使用単語辞書に登録された登録語と、前記文書解析部が抽出した前記単語情報に含まれる複数の単語の各々との文字列の類似性に基づき、前記類似性が近いほど値が高くなる指標である請求項 3 に記載の文書分析装置。

【請求項 5】

前記誤字脱字指標は、単語間の文字列の類似性に基づき、前記単語間の文字列の類似性が近いほど、値が高くなる指標である請求項 1 乃至 4 のいずれか 1 項に記載の文書分析装置。

【請求項 6】

前記誤字脱字指標は、前記文書における前記単語の出現頻度の偏りに基づき、前記文書における前記出現頻度の偏りが大きいほど、値が高くなる指標である請求項 1 又は 5 のいずれか 1 項に記載の文書分析装置。

10

20

30

40

50

【請求項 7】

前記パターン情報は、意味が類似しない少なくとも一文字違いの熟語の組み合わせを含み、複数の名詞又は未知語を結合した複合語において前記一文字違いの熟語が異なるだけの組み合わせたパターンを含む付記 1 乃至 6 のいずれか 1 項に記載の文書分析装置。

【請求項 8】

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を含むサーバ装置と、

少なくとも 1 つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する分析部を含む文書分析装置と、
を含む文書分析システム。

10

【請求項 9】

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶し、

少なくとも 1 つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する文書分析方法。

【請求項 10】

20

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶する処理と、

少なくとも 1 つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する処理と、
をコンピュータに実行させるプログラム。

【発明の詳細な説明】

【技術分野】

【0001】

30

本発明は、文書分析装置、分析システム、文書分析方法およびプログラムに関し、自然言語で書かれた文書から、仕様書など文書に固有の誤字・脱字を抽出する文書分析装置、文書分析システム、文書分析方法およびプログラムに関する。

【背景技術】

【0002】

近年、情報処理装置を用いて、自然言語で書かれた文書を分析して、その文書の誤字脱字を抽出するシステムが開発されている。

【0003】

特許文献 1 には、日本語の文書中の誤字を抽出する文書誤字抽出方式が開示されている。当該文書誤字抽出方式は、単語間の接続コストを定義するコストテーブルを利用して、誤字候補を出力する。

40

【0004】

特許文献 2 には、誤字脱字などの表記ゆれを含む文書に対し、高度な形態素解析を行い、当該形態素解析により得られた単語列データを用いて誤字脱字を修正する誤字脱字対応テキスト解析装置が開示されている。当該誤字脱字対応テキスト解析装置は、ある文字が別の文字に混同される程度（混同しやすさ）を数値化した文字混同確率を用いて、誤字脱字の修正候補を求める。

【先行技術文献】

【特許文献】

【0005】

50

【特許文献1】特開平4-213155号公報

【特許文献2】特許第5203324号公報

【発明の概要】

【発明が解決しようとする課題】

【0006】

特許文献1に記載の文書誤字抽出方式は、単語の品詞およびその他の情報を、左向きの構文属性（かかれ方）および右向きの構文属性（かかり方）として定義し、さらに各構文属性間のつながりやすさ（表層の並びの出現確率）を接続コストとして定義している。すなわち、当該文書誤字抽出方式は、各単語の構文属性のつながりやすさ、すなわち各単語の使われ方が文法的に正しいか否かという基準で誤字候補を検出する。

10

【0007】

また、特許文献2に記載の誤字脱字対応テキスト解析装置は、ある文字が、別の文字と同音の文字である場合や、別の文字と形が似ていて文字種が異なる文字である場合などに、文字混同確率の値が大きくなるように設定されている。すなわち、当該誤字脱字対応テキスト解析装置は、文字そのものが混同されやすいか否かという基準で誤字脱字を検出する。

【0008】

ここで、ある文字列が、その文字列だけに着目すると誤字脱字に該当しないが、特定の文書で用いられる場合には誤字脱字に該当することがある。例えば、システム構築に関して記載された文書において、「入荷金額」という言葉が正式に定義されているとする。この場合、「入荷額」はその文字列だけに着目すると誤字脱字とは言えないが、当該文書中においては「入荷金額」と記載されるべき文字列であり、誤字脱字に該当する。

20

【0009】

このような文書の内容に依存した誤字脱字は、単語の使われ方が文法的に正しいか否かという基準（特許文献1）や、文字が混同されやすいか否かという基準（特許文献2）で検出しようとしても、検出することができない。すなわち、特許文献1又は2に記載の技術では、文書の内容に依存した誤字脱字を抽出することができないという課題がある。

【0010】

本発明の目的は、上記課題に鑑み、文書の内容に依存する誤字脱字を抽出することができる文書分析装置、文書分析システム、文書分析方法及びプログラムを提供することにある。

30

【課題を解決するための手段】

【0011】

本発明の文書分析装置は、誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶する記憶部と、少なくとも1つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する分析部と、を含む。

【0012】

本発明の文書分析システムは、誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を含むサーバ装置と、少なくとも1つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する分析部を含む文書分析装置と、を含む。

40

【0013】

本発明の文書分析方法は、誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶し、少なくとも1つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報

50

に基づいて、誤字脱字に関する情報を生成する。

【0014】

本発明のプログラムは、誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶する処理と、少なくとも1つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する処理と、をコンピュータに実行させる。

【発明の効果】

【0015】

本発明は、文書分析装置、文書分析システム、文書分析方法及びプログラムにおいて、文書の内容に依存する誤字脱字を抽出することができるという効果がある。

【図面の簡単な説明】

【0016】

【図1】本発明の第1の実施形態における文書分析装置100の構成を示すブロック図である。

【図2】本発明の第1の実施形態における文書分析装置100の構成を実現するコンピュータの一例を示すブロック図である。

【図3】本発明の第1の実施形態における文書分析装置100の動作を示すシーケンス図である。

【図4】本発明の第2の実施形態における文書分析システム300の構成を示すブロック図である。

【図5】本発明の第2の実施形態における単語情報表の具体例を示す図である。

【図6】本発明の第2の実施形態におけるパターン情報の具体例を示す図である。

【図7】本発明の第2の実施形態におけるパターン情報の具体例を示す図である。

【図8】本発明の第2の実施形態におけるノイズ情報の具体例を示す図である。

【図9】本発明の第2の実施形態における誤字脱字に関する情報の具体例を示す図である。

【図10】本発明の第2の実施形態における誤字脱字に関する情報の具体例を示す図である。

【図11】本発明の第3の実施形態における文書分析装置1000の構成を示すブロック図である。

【発明を実施するための形態】

【0017】

まず、本発明の実施形態を理解し易くするために、特許文献1及び2に記載された技術について説明し、本発明の実施形態において解決すべき課題を明らかにする。

【0018】

特許文献1に記載された文書誤字抽出方式は、機械処理用に構成された単語辞書と、単語間の接続コストを定義するコストテーブルと、同音語をグループ化して互いに参照可能にした同音語辞書と、を備える。さらに、文書誤字抽出方式は、形態素解析手段と、同音語候補チェック手段と、誤字候補出力手段とを備える。

【0019】

形態素解析手段は、単語辞書とコストテーブルを参照して入力文書中の隣接単語間の接続コストを求める。同音語候補チェック手段は、単語辞書、コストテーブルおよび同音語辞書を参照して、接続コストが閾値より低い同音語をチェックする。誤字候補出力手段は、チェック結果に基づいて誤字候補を出力する。

【0020】

このような構成を有する文書誤字抽出方式は、具体的に次のように動作する。すなわち、形態素解析手段が、単語辞書の辞書情報に従って入力文書を単語単位に切り分ける。さらに、形態素解析手段は、コストテーブルを参照し、表層の並びの出現確率に基づき、隣

10

20

30

40

50

接单語間の接続コストを求める。次に同音語候補チェック手段が、形態素解析の結果から単語辞書に登録がなかった未登録語および接続コストがあらかじめ定められた閾値より高い単語に対して、単語辞書、コストテーブルおよび同音語辞書を参照して接続コストが閾値より低い同音語があるかどうかをチェックする。最後に誤字候補出力手段は、チェックされた接続コストが閾値より低い同音語があるとされた単語を誤字候補として出力する。

【0021】

以上のように、特許文献1に記載された文書誤字抽出方式は、単語の品詞その他の情報から表層の並びの出現確率を定義したコストテーブルを利用して、文法的に、より確からしい単語の並びが存在する場合に限って誤字候補を出力する。これにより、人間が目チェックする場合に比べて、利用者の負担なしで間違いの少ない誤字の抽出を行うことができる。

10

【0022】

一方、特許文献2のテキスト解析装置は、辞書登録語での出現文字と入力文での出現文字で変化があった箇所の状況である近似文字状況毎に、ある文字が別の文字に混同される程度（混同しやすさ）を数値化した文字混同確率を格納した文字混同確率記憶手段を含む。さらに、テキスト解析装置は、ある読みに対する漢字を格納した読み漢字対応記憶手段と、統計的言語モデルを格納した統計的言語モデル記憶手段と、入力テキストを形態素解析して単語列データを出力する形態素解析手段とを含む。また、テキスト解析装置は、入力テキストの単語の所定の文字長の単語について近似照合し、近似辞書照合単語列データを出力する近似辞書照合手段と、単語列データと近似辞書照合単語列データを用いて誤字脱字修正を行う誤字脱字修正手段とを含む。

20

【0023】

特許文献2のテキスト解析装置は、形態素解析手段が入力文を形態素解析して使われている単語の表記や品詞といった単語列データを出力する。さらに、近似辞書照合手段が、同じ入力文に対して近似辞書照合処理を実行して元キーワード、照合開始位置、照合終了位置、誤り開始位置、誤り文字数、一致タイプといった近似辞書照合単語列データを出力する。次に、誤字脱字修正手段は、文字混同確率記憶手段、読み漢字対応記憶手段、統計的言語モデル記憶手段にそれぞれ格納された、ある文字が別の文字に混同されやすさを数値化した文字混同確率と、同一の読みを持つ漢字と、統計的言語モデルとを用いて、次の処理を実行する。すなわち、誤字脱字修正手段は、単語列データおよび近似辞書照合単語列データを受け、それらに存在する単語について、その照合による一致タイプと近似文字状況の2つの観点から、各単語に重み付けを行い、重みつき単語列データを作成する。さらに、誤字脱字修正手段は、重みつき単語列データを用いて、統計的言語モデル記憶手段に格納された統計的言語モデルに基づき、統計的言語モデルを参照して単語列毎の重みを考慮した表記列と品詞列の同時確率を最大化するような単語分割（最尤単語列）に、単語列データの文字列を変換する。この変換によって、誤字脱字が修正される。

30

【0024】

このような構成により、誤字脱字が含まれるテキストであっても入力文中の単語を認定して品詞などの辞書情報を付与する形態素解析を正しく行うことができる。

【0025】

40

しかしながら、特許文献1及び2には、自然言語で書かれた文書によって誤字脱字の量又はレベルが違うという点に着目していない。その「レベル」とは、「文書における誤字脱字への配慮の程度」を表すものとする。したがって、特許文献1及び2には、その量又はレベルに基づいて誤字脱字を分析し、それを検出することができないという課題がある。本発明の実施形態は、当該課題を解決する。自然言語で書かれた文書には、例えば、システム構築における仕様書がある。仕様書は、開発や実装作業を行うためのものであり、使用可能な文字が厳格であり、誤字脱字が少ない文で構成されなければならない。また、一般的な文書の場合、誤字脱字のレベルが低い場合もある。このように、文書における誤字脱字の量又はレベルに応じた誤字脱字の抽出が必要になる。

【0026】

50

＜実施形態 1＞

次に、本発明の実施形態について、図面を参照して詳細に説明する。図 1 は、本発明の第 1 の実施形態における文書分析装置 100 の構成を示すブロック図である。

【0027】

図 1 を用いて、本発明の第 1 の実施形態における文書分析装置 100 について説明する。なお、図 1 に付記した図面参照符号は、理解を助けるための一例として各要素に便宜上付記したものであり、本発明に対するなんらの限定を意図するものではない。

【0028】

図 1 に示すように、本発明の第 1 の実施形態における文書分析装置 100 は、入力部 10、分析部 20、出力部 30、及びノイズデータベース 40 を含む。分析部 20 は、文書解析部 21、誤字脱字指標生成部 22、ノイズ情報検出部 23、誤字脱字抽出条件変更部 24、及び誤字脱字抽出部 25 を含む。また、文書分析装置 100 は、電子機器内もしくはサーバと電子機器およびこれらを相互に接続するインターネット等の情報通信ネットワークからなるシステム等で構成されてもよい。

【0029】

図 1 に示す文書分析装置 100 は、例えば、自然言語で記載された文書に対して、分析対象の文書が誤字脱字にどの程度配慮して書かれているかという配慮程度を情報として利用する。文書分析装置 100 は、その情報に基づいて、誤字脱字の可能性のある表現から本当に誤字脱字とみなすべき表現を絞り込む際の条件を変更し、変更された条件を使って文書に固有の誤字脱字を抽出する。

【0030】

ここで、文書分析装置 100 は、電子機器で構成される場合、例えば、プログラム制御により動作するコンピュータで実現可能である。図 2 は、本発明の第 1 の実施形態における文書分析装置 100 の代表的なハードウェア構成図である。図 2 に示すように、文書分析装置 100 は、例えば CPU (Central Processing Unit) 1、RAM (Random Access Memory) 2、記憶装置 3、通信インターフェース 4、入力装置 5、出力装置 6 を含む。

【0031】

図 1 における分析部 20、文書解析部 21、誤字脱字指標生成部 22、ノイズ情報検出部 23、誤字脱字抽出条件変更部 24、及び誤字脱字抽出部 25 は、プログラムを RAM 2 に読み出し、実行する CPU 1 によって実現される。入力部 10 及び出力部 30 が情報の送受信を行う動作は、例えば CPU 1 の OS (Operating System) を実行することで実現される。ノイズデータベース 40 は、例えばハードディスクやフラッシュメモリ、光ディスクなどのディスクメモリのほか、半導体メモリを含む。

【0032】

通信インターフェース 4 は、CPU 1 に接続され、ネットワーク或いは外部記憶媒体に接続される。外部データが通信インターフェース 4 を介して CPU 1 に取り込まれても良い。入力装置 5 は、例えばキーボードやマウス、タッチパネルである。出力装置 6 は、例えばディスプレイである。

【0033】

また、文書分析装置 100 は、例えば、データを入力する入力装置と、データ処理装置と、データ処理装置での処理結果を出力する出力装置と、種々のデータベースとして働く補助記憶装置とで構成されていても良い。その場合、データ処理装置は、プログラムを記憶するリードオンリメモリ (ROM) と、データを一時的に記憶するワークエリアとして使用されるランダムアクセスメモリ (RAM) と、ROM に記憶されたプログラムに従って、RAM に記憶されているデータを処理する中央処理装置 (CPU) とから構成される。

【0034】

この場合、データ処理装置が入力部 10、文書解析部 21、誤字脱字 22、ノイズ情報検出部 23、誤字脱字抽出条件変更部 24、誤字脱字抽出部 25 として動作する。補助記

10

20

30

40

50

憶装置がノイズデータベース40として動作する。出力装置が出力部30として動作する。

【0035】

なお、図2に示すハードウェア構成は、一例にすぎず、図1に示す各部それぞれが独立した論理回路で構成されていても良い。

【0036】

次に、文書分析装置100を構成する各構成要素について説明する。

【0037】

入力部10は、文書もしくは文書群（以下、少なくとも1つの文書と記載する）を受け付け、文書解析部21に出力する。少なくとも1つの文書とは、例えば、文書に固有の誤字脱字を含む可能性が有り、修正が必要な誤字脱字の箇所を分析する必要がある文書である。また、少なくとも1つの文書は、例えば、自然言語で構成された文書である。

【0038】

文書解析部21は、入力部10から出力された少なくとも1つの文書を構成する各文章に対して形態素解析を適用し、各文章または各文に使用されている全単語の情報（以下、単語情報と記載する）を抽出する。ここで、文書解析部21は、例えば、形態素解析した各文章または各文の全単語について、自立語と付属語とを個別の単語（以下、1単語と記載する）の単語情報として抽出する。自立語とは、名詞、動詞、形容詞などの、単独で意味をなす語である。付属語とは、助詞などの、単独で意味をなさない語である。

【0039】

また、文書解析部21は、名詞や未知語が助詞などを挟まずに隣接している場合、例えば、隣接する名詞および未知語同士で結合されてなる複合語を1単語として抽出する。また、複合語は、例えば、全ての隣接語同士で結合された語が1単語の単語情報であってもよい。ここで、未知語とは、文書解析部21が少なくとも1つの文書を形態素解析するために必要な辞書、に登録されていない語である。

【0040】

また、文書解析部21は、同一の文字列からなる単語であっても文書内における出現箇所が異なれば、それぞれについて単語情報の抽出を行う。さらに、上述の単語情報は、例えば、使用されている単語の文字列と、単語の文書内での存在位置に関する情報を含む。単語の文書内での存在位置に関する情報とは、使用箇所が同定可能な情報である。単語の文書内での存在位置に関する情報は、単語の存在する文の出現順位や頁、目次上の章や節、項などである。

【0041】

誤字脱字指標生成部22は、文書解析部21が抽出した単語情報を用いて、誤字脱字指標を算出する。誤字脱字指標とは、文書解析部21で検出された文書内の全単語を組み合わせる合成語または複合語のそれぞれに対して、誤字脱字の可能性を示す値である。誤字脱字指標は、例えば、文書が有する合成語や複合語が、その文書においてどの程度誤字脱字と判断できるかを示す誤字度（誤字レベル）、または誤字の確率であって、同じ合成語や複合語であっても文書の種類、ジャンルあるいは内容の厳格性に依りて異なる値を有する。

【0042】

ここで、誤字脱字指標は、例えば、単語間の文字列の類似性及び文書中の各単語の出現頻度の偏りに基づき、単語間の文字列の類似性がより近い単語の組み合わせ又は文書中の単語間の出現頻度の偏りがより大きい単語の組み合わせほど、値がより高くなることが望ましい。それは、当該値が高くなることは、対象の単語が誤字脱字である可能性が高いことを示しているからである。また誤字脱字指標は、例えば、使用単語辞書に登録された登録語と文書解析部21で抽出された各単語の文字列の類似性がより近いほど値がより高くなることが望ましい。それは、当該値が高くなることは、対象の単語が誤字脱字である可能性が高いことを示しているからである。使用単語辞書とは、事前に文書に使用可能な単語の一覧を登録した辞書（記憶デバイス）である。この場合、単語の組み合わせは文書内

10

20

30

40

50

の単語の組み合わせではなく、登録語と抽出された各単語との組み合わせとなる。文書解析部 21 は、使用単語辞書を含んでもよい。

【0043】

また、単語間の文字列の類似性を定量化する指標としては、編集距離と単調減少の関係にある指標が適している。編集距離とは、2つの文字列がどれくらい似ているか（又は異なっているか）を表す指標（数）である。例えば、「入荷額」と「入荷金額」の編集距離は1である。また、例えば、「対処」と「対象」の編集距離は1である。

【0044】

分析部 20 が少なくとも1つの文書から、誤字脱字を含む文字または単語を、誤字脱字指標に基づいて検出する場合に、ノイズデータベース 40 は、その検出のために必要な情報を蓄積する。すなわち、ノイズデータベース 40 は、誤検出となる文字または単語のパターン（誤検出パターン）を抽出可能にする情報（以下、パターン情報と記載する）を、事前に収集し、蓄積する。ノイズデータベース 40 は、ノイズ情報検出部 23 からの問い合わせ、すなわち、文書解析部 21 で検出された、当該文書の任意の単語の組み合わせ（例えば、合成語や複合語）が誤検出パターンに該当するかどうかに関する問い合わせを受ける。さらに、ノイズデータベース 40 は、その問い合わせに対し、問い合わせ対象の単語の組み合わせが誤検出パターンと一致するパターン情報を検索し、応答する。

【0045】

ここで、問い合わせ対象の単語の組み合わせに関する誤字脱字指標が、例えば、単語間の文字列の類似性と、文書中の各単語の出現頻度の偏りに基づく場合、誤検出となるパターンとしては、単語間の文字列の類似性が近くなる接辞語が一方に付随するパターンが考えられる。この場合の誤検出となるパターンを抽出可能にする情報は、接辞語の語群が該当する。

【0046】

また、ノイズデータベース 40 は、文書の種類、ジャンル、及び誤字脱字の量等の、文書の内容に依存する固有の誤字脱字を検出する場合において、例えば、文書の内容ごとに異なる誤検出パターンのパターン情報を蓄積してもよい。ノイズ情報検出部 23 が当該ノイズデータベース 40 に対して問い合わせをした場合、ユーザが、入力装置 5 から文書の種類やジャンル等を決定してから、ノイズデータベース 40 は、各文書に対するパターン情報を検索してもよい。

【0047】

他にも、誤字脱字指標が単語間の文字列の類似性と文書中の各単語の出現頻度の偏りに基づいて算出された場合、誤検出となるパターンは、例えば、複数の名詞や未知語を結合した複合語で一文字違いの熟語が異なるだけの組み合わせたパターンである。この誤検出パターンをもつ問い合わせ対象の単語の組み合わせを、ノイズデータベース 40 から抽出可能にするパターン情報として、意味が類似しない一文字違いの熟語の組み合わせが適当である。また、誤字脱字指標が使用単語辞書に登録された登録語と抽出された各単語の文字列の類似性とに基づく場合、誤検出となるパターンは、例えば、抽出された各単語に類似する文字列があるとした登録語、とは別に完全一致する登録語があるとするパターンである。この場合の誤検出となるパターンを抽出可能にする情報は、抽出された各単語と別の登録語との一致状況が該当する。

【0048】

なお、ノイズデータベース 40 は、例えば、情報処理装置 100 の外部に存在し、ネットワーク上にあるデータベースであっても良い。また、その場合は、文書分析装置 100 とノイズデータベース 40 は、ネットワークで接続されている。

【0049】

ノイズ情報検出部 23 は、誤字脱字指標生成部 22 で算出した誤字脱字指標に基づき、各単語の組み合わせに対して、誤字脱字指標で誤検出となるパターンの有無をノイズデータベース 40 に問合せる。さらに、ノイズ情報検出部 23 は、その問い合わせに対する応答としてノイズとなる組み合わせを受け、それをノイズ情報として検出する。

【0050】

ここで、誤字脱字指標が単語間の文字列の類似性と文書中の各単語の出現頻度の偏りに基づく場合において、検出されるノイズ情報は、誤字脱字指標で誤検出の主因となる情報である。その情報は、単語の組み合わせの一方に接辞語が付いた組み合わせや、複数の名詞や未知語を結合した複合語であって一文字違いの熟語が異なるだけの組み合わせである。

【0051】

また、他の誤字脱字指標は、使用単語辞書に登録された登録語と、抽出された各単語が類似する文字列があるとした登録語とは別に完全一致する登録語があるパターンでの抽出された各単語と、文字列が類似する登録語の組み合わせなどが考えられる。

10

【0052】

誤字脱字抽出条件変更部24は、ノイズ情報を利用して所定の誤字脱字抽出条件を変更するための変更ルールに基づき、誤字脱字語（誤字脱字を含む文字または単語）の抽出条件を変更する。ここで、誤字脱字抽出条件を変更するための変更ルール（以降、誤字脱字抽出条件変更ルールとも言う）は、例えば、誤字脱字指標の値が高い（大きい）組み合わせ群におけるノイズ情報の割合が高い文書ほど、誤字脱字を検出するための誤字脱字抽出条件を厳しくするルールであってもよい。その理由として、誤字脱字指標が高い組み合わせ群におけるノイズ情報の割合が高い文書は、誤字脱字に配慮して書かれた文書であり、誤字脱字の数が少ないと推定できるからである。

【0053】

20

誤字脱字抽出条件変更ルールは、例えば、算出した誤字脱字指標に基づき、各単語の組み合わせを誤字脱字指標の高い順に並べたリストの上位の任意の数の組み合わせを抽出するというルール（以下、変更ルール1と記載する）がある。また、誤字脱字抽出条件変更ルールは、例えば、誤字脱字指標が一定値以上の組み合わせに対して、ノイズ情報に該当した組み合わせの割合と単調減少の関係にある指標に基づき、抽出する組み合わせの数を変更させるというルール（以下、変更ルール2と記載する）である。また、誤字脱字抽出条件変更ルールは、例えば、ノイズ情報に該当した組み合わせの割合と単調増加の関係にある指標に基づき、抽出する組み合わせの誤字脱字指標の下限値を変更するというルール（以下、変更ルール3と記載する）である。

【0054】

30

誤字脱字抽出部25は、誤字脱字指標生成部22で、誤字脱字指標を算出した文書内の全ての単語の組み合わせから、ノイズ情報検出部23でノイズ情報に該当した誤検出の組み合わせを除外する。さらに、誤字脱字抽出部25は、誤字脱字抽出条件変更部24で誤字脱字抽出条件変更ルールによって変更した誤字脱字語の抽出条件に基づき、誤字脱字語を抽出する。誤字脱字抽出条件変更ルールが、変更ルール1の場合は、上位の組み合わせから、変更した組み合わせ数分の組み合わせを抽出し、各組み合わせ中で出現頻度が低い方の単語を誤字脱字語として抽出する。同様に、誤字脱字抽出条件変更ルールが、変更ルール2の場合も、上位の組み合わせから、変更した組み合わせ数分の組み合わせを抽出し、各組み合わせ中で出現頻度が低い方の単語を誤字脱字語として抽出する。

【0055】

40

一方、誤字脱字抽出条件変更ルールが、変更ルール3である場合は、上位の組み合わせから、変更した下限値までの組み合わせを抽出し、各組み合わせ中で出現頻度が低い方の単語を誤字脱字語として抽出する。また、誤字脱字抽出部25は、例えば、少なくとも抽出した各誤字脱字語、当該誤字脱字語に対応する誤字脱字指標の値および正しいと推定される単語を誤字脱字に関する情報として抽出してもよい。ここで、誤字脱字語に関する情報は、例えば、少なくとも文書が有する正しい単語と、正しい単語に対応する誤字脱字を含む単語と、誤字脱字指標との対応関係（例えば、リスト）を含んでもよい。

【0056】

なお、誤字脱字指標生成部22で誤字脱字指標を算出した単語の組み合わせが使用単語辞書に登録された登録語と抽出された各単語の組み合わせの場合、誤字脱字抽出部25は

50

、抽出された単語を誤字脱字語として抽出する。

【0057】

出力部30は、誤字脱字抽出部25で抽出した各誤字脱字語について、対応する誤字脱字指標の値および文書内での存在位置、正しいと推定される単語等を誤字脱字に関する情報として出力する。ここで、出力部30が出力する誤字脱字に関する情報は、上述の対応関係（例えば、リスト）を含んでもよい。また、出力部30が出力する誤字脱字に関する情報は、上記内容に限定されない。出力部30の出力形態は、所要の形態で出力しても良い。出力部30は、例えば、文書内における各誤字脱字を色分けや太字による強調、文字の拡大などで明示することで、文書全体を出力するなどである。

【0058】

また、出力部30の出力形態としては、各誤字脱字語を抽出した表などの形態であっても良い。また、出力部30の出力形態としては、誤字脱字指標の値によって色分けや太字による強調もしくは単語の文字の大きさなどに強弱を与えるなどしても良い。また、出力部30の出力形態としては、誤字脱字毎の誤字脱字指標の値を文書全体または任意の範囲で集計し、文書の品質を表す指標として表形式で出力する形態であっても良い。また、各出力形態を選択できるようにして、ベースとなる表示形態から必要に応じて表に移行できるようにしても良い。

【0059】

次に、図1の構成図及び図3のシーケンス図を参照して、本発明の第1の実施形態における文書分析装置100の全体の動作について詳細に説明する。なお、図3に示すシーケンス図及び以下の説明は処理例であり、適宜求める処理に応じて処理順等を入れ替えたり処理を戻したり繰り返したりすることを行っても良い。

【0060】

図3に示すように、入力部10は、少なくとも1つの文書の入力を受け付け（文書を読み込み）、文書解析部21に出力する（ステップS1）。ここで、少なくとも1つの文書とは、固有の誤字脱字を含む可能性が有り、修正が必要な誤字脱字の箇所を分析する文書もしくは文書群を示す。

【0061】

文書解析部21は、入力された少なくとも1つの文書を構成する各文章に対して形態素解析を適用することで、各文章（少なくとも1つの文書）内に含まれている単語情報の抽出を行う（ステップS2）。また、文書解析部21は、例えば、形態素解析機能を備え、少なくとも1つの文書に対して形態素解析を行い、単語情報を抽出しても良い。

【0062】

誤字脱字指標生成部22は、文書解析部21が抽出した単語情報を用いて、文書内の全単語を組み合わせ、当該全単語の組み合わせに対して誤字脱字指標を算出する（ステップS3）。ここで、誤字脱字指標とは、誤字脱字の可能性を示す指標（数）である。

【0063】

ノイズデータベース40は、文書に固有の誤字脱字を誤字脱字指標で検出しようとした場合に誤検出となるパターンを抽出可能にする情報を、事前に収集して蓄積する。また、ノイズデータベース40は、任意の単語の組み合わせが誤検出パターンに該当するかどうかに関する問い合わせに対し、問い合わせ対象の単語の組み合わせが誤検出パターンと一致するか検索し、応答する（ステップS4）。

【0064】

ノイズ情報検出部23は、誤字脱字指標生成部22で算出した誤字脱字指標に基づき、各単語の組み合わせに対して、誤字脱字指標で誤検出となるパターンの有無をノイズデータベース40に問合せ。ノイズ情報検出部23は、ノイズデータベース40からの応答としてノイズとなる組み合わせを受け取り、それをノイズ情報として検出する（ステップS5）。

【0065】

誤字脱字抽出条件変更部24は、ノイズ情報を利用して所定の誤字脱字抽出条件変更ル

10

20

30

40

50

ールに基づき、誤字脱字語の抽出条件を変更する（ステップS6）。

【0066】

誤字脱字抽出部25は、誤字脱字指標生成部22で、誤字脱字指標を算出した文書内の全ての単語の組み合わせから、ノイズ情報検出部23でノイズ情報に該当した組み合わせを除外する。そして、誤字脱字抽出部25は、さらに誤字脱字抽出条件変更部24で誤字脱字抽出条件変更ルールによって変更した誤字脱字語を抽出するための抽出条件に基づき、誤字脱字語を抽出する（ステップS7）。

【0067】

出力部30は、誤字脱字抽出部25で抽出した各誤字脱字語について、対応する誤字脱字指標の値および文書内での存在位置、正しいと推定される単語を誤字脱字情報として出力する（ステップS8）。

【0068】

上記のとおり、本発明の第1の実施形態における文書分析装置100において、入力部10が少なくとも1つの文書を受け付け、文書解析部21が入力部10から出力された少なくとも1つの文書を構成する各文章に対して形態素解析を適用し、各文章または各文に使用されている全単語及び複合語の情報を抽出する。文書分析装置100の誤字脱字指標生成部22が、単語情報を用いて、誤字脱字指標を算出し、ノイズ情報検出部23が、誤字脱字指標に基づき、各単語の組み合わせに対して、誤字脱字指標で誤検出となるパターンの有無をノイズデータベース40に問合せ。さらに、文書分析装置100のノイズ情報検出部23は、その問い合わせに対する応答としてノイズとなる組み合わせを受け、それをノイズ情報として検出する。

【0069】

文書分析装置100の誤字脱字抽出条件変更部24は、ノイズ情報を利用して所定の誤字脱字抽出条件を変更するための変更ルールに基づき、誤字脱字語の抽出条件を変更する。文書分析装置100の誤字脱字抽出部25は、文書内の全ての単語の組み合わせから、ノイズ情報に該当した誤検出の組み合わせを除外し、誤字脱字抽出条件変更ルールによって変更された誤字脱字語の抽出条件に基づき、誤字脱字語を抽出する。文書分析装置100の出力部30は、誤字脱字抽出部25で抽出した各誤字脱字語について、対応する誤字脱字指標の値および文書内での存在位置、正しいと推定される単語等を誤字脱字に関する情報として出力する。したがって、当該文書分析装置100は、文書の内容に依存する誤字脱字を抽出することができる。

【0070】

また、文書分析装置100は、自然言語で書かれた文書に固有の誤字脱字を含む文書から、誤字脱字の可能性の高い表現に限定した抽出、優先的な修正が必要な箇所限定した指摘、文書の修正時の負荷の低減、及び文書作成や文書レビューの効率化を可能とする。

【0071】

また、文書分析装置100は、システム開発における独自の語の定義を持つ要件定義書や設計書等の、スペルチェッカーでは検出できない誤字脱字を含む文書に関して、誤字脱字に配慮されている文書かどうかで提示する誤字脱字の可能性のある語を絞り込む。それにより、文書分析装置100は、効率的に文書の修正ができる。また、文書分析装置100は、読み手の不安や不信感が起きる状況などが減少し、円滑なコミュニケーションや顧客満足の向上などシステム開発の効率化に関する用途に適用できる。

【0072】

また、文書分析装置100は、誤字脱字を含む文書の誤字脱字を抽出する際に、誤字脱字に気を使って書かれている文書であれば、提示する誤字脱字の候補を絞り込むという判定基準を変更できる。また、文書分析装置100は、誤字脱字を確認して修正する作業を効率化することが可能になる。その理由としては、文書分析装置100が、文書において誤字脱字が文書内で使用されている用例を分析し、誤字脱字でない用例をノイズ情報として、ノイズ情報の多い文書ほど、誤字脱字として出力する候補とみなす基準を厳しくするように構成しているためである。

10

20

30

40

50

【0073】

なお、上記本発明の実施形態における文書分析装置100は、文書分析方法として実現され得る。また、上記本発明の実施形態における文書分析装置100は、プログラムによりコンピュータによって実行させるようにしても良い。

【0074】

＜第2の実施形態＞

本発明を実施するための第2の形態における文書分析システム300の構成について、図面を参照して説明する。なお、本発明の第2の実施形態において、本発明の第1の実施形態と同様の構成については、説明を省略する。

【0075】

図4は、本発明の第2の実施形態における文書分析システム300の構成を示す図である。図4に示すように、文書分析システム300は、文書分析装置100'とサーバ装置110（例えば、イントラネットサーバZ）と、を含む。文書分析装置100'は、入力部10と、分析部20'と、出力部30と、を含む。図4乃至図10を参照して、具体的な例を用いて、本発明を実施するための第2の形態における文書分析システム300の動作について説明する。

【0076】

まず、文書分析システム300は、情報システム構築に関する提案書や仕様書といった、誤字脱字を排除すべき文書（例えば、文書D）内で、誤字脱字を含む単語（例えば、誤字脱字語Wa）について、単語が誤字脱字である可能性を示す誤字脱字指標（例えば、誤字度IA）を各誤字脱字語Waについて算出する。文書分析システム300は、各誤字脱字語Waの文書中の位置や誤字度IA、正しいと推定される単語Warを基に誤字脱字に関する情報（例えば、誤字脱字情報A）を抽出し、出力する。

【0077】

ここで、分析部20'は、上述した本発明の第1の実施形態における文書分析装置100の文書解析部21、誤字脱字指標生成部22、ノイズ情報検出部23、誤字脱字抽出条件変更部24、及び誤字脱字抽出部25と同様の機能を有する。サーバ装置110は、上述した本発明の第1の実施形態における文書分析装置100のノイズデータベース40と同様の構成と機能を有する。

【0078】

また、分析部20'は、例えば、文書を解析するシステム（例えば、文書解析システムY）として、文書を分析するユーザ（例えば、分析実施者K）の持つPC（Personal Computer：パーソナルコンピュータ）端末上で動作してもよい。その場合、文書解析システムYは、入力部10及び出力部30を介して、分析実施者Kが誤字脱字情報Aを抽出したい文書群を構成する文章の入力と、誤字脱字情報Aの出力を実現する。また、文書分析システムYは、分析部20'が有する構成及び機能を含む。

【0079】

イントラネットサーバZは、通信ネットワークを介して文書解析システムYを実装した分析実施者Kの持つPC端末と接続されている。イントラネットサーバZは、文書解析システムYからの任意の単語の組み合わせが誤検出パターンに該当するかどうかに関する問い合わせに対し、任意の単語の組み合わせに関する誤検出パターンの検索を可能にする装置である。

【0080】

また、入力部10は、例えば、PC端末の入力部として動作してもよい。また、出力部30は、例えば、PC端末の出力部として動作してもよい。

【0081】

次に、本発明の第2の実施形態における文書分析システム300の動作について説明する。

【0082】

入力部10は、情報システム構築に関する提案書や仕様書といった、分析実施者Kが誤

10

20

30

40

50

字脱字を排除するための誤字脱字情報 A を得たい文書 D の入力を受け付ける。そして、分析部 20' は、文書 D を構成する文章または文毎に形態素解析を適用し、文書 D に含まれる全ての単語 W_i ($i = 1, 2, \dots, n$, n は自然数) について単語の文字列を単語情報として抽出する。この動作は、上述した本発明の第 1 の実施形態における文書分析装置 100 の文書解析部 21 に相当する。

【0083】

ここで、名詞や辞書に登録されていない未知語が助詞などを挟まずに隣接している場合は、隣接する名詞および未知語を全て結合した複合語を単語 W とする。さらに、分析部 20' は、文書内の文の出現順に通し番号として文番号を付け、各単語を含む文の文番号を単語情報に加える。

【0084】

さらに分析部 20' は、文書 D に含まれる全ての単語 W_i 間の誤字度 $I A$ を算出する。この動作は、上述した本発明の第 1 の実施形態における文書分析装置 100 の誤字脱字指標生成部 22 に相当する。例えば、単語 W_i と単語 W_j ($j = 1, 2, \dots, n$, n は自然数) 間の誤字度 $I A_{ij}$ は以下の式 (1) で算出される。

$$I A_{ij} = \alpha / \sqrt{(L_{ij})} + \beta \times B_{ij} \quad (1)$$

ここで、 α は単語間の文字列の類似性に関する重みづけ係数である。 L_{ij} は単語 W_i と単語 W_j 間の編集距離である。 $\sqrt{(L_{ij})}$ は、 $(L_{ij})^{1/2}$ である。 β は文書中の各単語の出現頻度の偏りに関する重みづけ係数である。 B_{ij} は単語 W_i と単語 W_j 間の出現頻度の偏りを示す指数である。偏り B_{ij} は、単語 W_i が m_i 個と単語 W_j が m_j 個間の以下の式 (2) で算出される。

$$B_{ij} = |m_i - m_j| / (m_i + m_j) \quad (2)$$

例えば、 W_i が「入荷額」及び W_j が「入荷金額」である場合は、 L_{ij} は 1 である。また、例えば、 W_i が 2 個及び W_j が 3 個である場合は、 $B_{ij} = |2 - 3| / (2 + 3)$ となり、 B_{ij} は 0.2 である。

【0085】

図 5 は、本発明の第 2 の実施形態における単語情報表の具体例を示す図である。図 5 に示すように、単語情報表とは、分析部 20' が文書から抽出した単語情報及び誤字脱字指標を含む表 (リスト) である。図 5 に示すように、単語情報表は、例えば、文書 D に含まれる一部の単語 W の間の誤字度 $I A$ を算出した結果を示す。単語情報表は、暫定順位、正しいと推定される単語 W_{ar} の候補、 W_{ar} の出現個数 $m_{W_{ar}}$ 、誤字脱字語 W_a の候補、 W_a の出現個数 m_{W_a} 、及び誤字度 $I A$ を含む。ここで、単語情報表は、例えば、図 5 に示すリストの内容に限定されず、分析部 20' が文書から抽出した単語情報及び誤字脱字指標を備えるために必要な情報であれば、どのような情報であってもよい。

【0086】

暫定順位とは、分析部 20' が単語 W に対して算出した誤字度 $I A$ の順位である。暫定順位は、例えば、誤字度 $I A$ の高い値から順に並べる。正しいと推定される単語 W_{ar} の候補とは、分析部 20' が取得した単語情報から推定される正しい単語 W_{ar} を示す単語群である。 W_{ar} の出現個数 m とは、文書 D 内に存在する正しいと推定される単語 W_{ar} のそれぞれが出現する個数を示す。誤字脱字語 W_a の候補とは、分析部 20' が取得した単語情報から得られる誤字脱字語 W_a を示す単語群である。 W_a の出現個数 m_{W_a} とは、文書 D 内に存在する誤字脱字語 W_a のそれぞれが出現する個数を示す。

【0087】

図 5 に示すように、単語情報表は、例えば、暫定順位が 1 位、正しいと推定される単語 W_{ar} の候補が「単語出現頻度」、 W_{ar} の出現個数 $m_{W_{ar}}$ が 46 個、誤字脱字語 W_a の候補が「単語発現頻度」、 W_a の出現個数 m_{W_a} が 1 個、及び誤字度 $I A$ が 0.98、を示す。また、単語情報表は、例えば、出現個数が多い単語 W を左側に、少ない単語 W を右側に配置し、誤字度 $I A$ が大きい順に並べて表記してもよい。

【0088】

ここで、図 5 に示すように、誤字度 $I A$ は、例えば 0 から 1 の間で値を取り、当該値が

大きいほど誤字脱字である可能性が高いことを示す。また、図5に示す単語情報表に記載される、出現個数が多い単語Wが「正しいと推定される単語W a rの候補」に、リストで右側に記載した、出現個数が少ない単語Wが「誤字脱字語W aの候補」に相当する。

【0089】

図6及び図7は、本発明の第2の実施形態におけるパターン情報の具体例を示す図である。図6及び図7は、サーバ装置110が有する、誤字脱字指標を分析部20'が文書から抽出した単語情報及び誤字脱字指標を含む単語情報表である。さらに、サーバ装置110（例えば、イントラネットサーバZ）は、誤字度IAの算出指標によって誤字脱字語W aを抽出する際に、誤検出となるパターンを抽出可能にするパターン情報（例えば、誤検出パターン情報C）を収集し、蓄積する。分析部200が上述した式（1）に基づく誤字度IAの算出方法を利用する場合、パターン情報は、例えば、接頭語や接尾語といった接辞語が一方に付随するパターンなどが考えられる。図6に示すように、誤検出パターン情報Cは、誤検出となるパターンを抽出可能にする、接頭語や接尾語を取集したリストである。

10

【0090】

また、上述した式（1）に基づく誤字度IAの算出方法を利用する場合、他の誤検出となるパターンとしては、「システム安全性」のような複数の名詞や未知語を結合した複合語に対して、「システム完全性」のように、複合語を構成する語が「安全」と「完全」といった一文字違いの熟語が異なるだけの組み合わせとなっているパターンが考えられる。図7に示すように、誤検出パターン情報Cは、意味が近くない一文字違いの熟語の組み合わせを取集したリストである。例えば、検出パターン情報Cは、「設置」と「設定」をひとまとまりとした情報を含む。

20

【0091】

さらにサーバ装置110は、任意の単語や表現の情報を抽出する検索エンジンなどの機能も提供することで、分析部20'からの問い合わせに応じて、問い合わせ対象の単語組み合わせに誤検出パターン情報Cに該当する単語組み合わせが存在するかどうかを判定し、存在する場合はその単語組み合わせを誤検出組み合わせとして提示する。この動作は、上述した本発明の第1の実施形態における文書分析装置100のノイズデータベース40に相当する。

【0092】

次に、分析部20'は、誤字度IAを算出した「正しいと推定される単語W a rの候補」と「誤字脱字語W aの候補」の組み合わせに対して、サーバ装置110に問い合わせ、パターン情報に該当する組み合わせパターンをノイズ情報（例えば、ノイズ情報N）として抽出する。この動作は、上述した本発明の第1の実施形態における文書分析装置100のノイズ情報検出部23に相当する。

30

【0093】

図8は、本発明の第2の実施形態におけるノイズ情報の具体例を示す図である。図8に示すように、分析部20'は、例えば、図5に示す「正しいと推定される単語W a rの候補」と「誤字脱字語W aの候補」の組み合わせから、図8に示す「集積地区」と「各集積地区」、「対象エリア」と「元対象エリア」などの複合語パターン、および、図8に示す「設置担当者」と「設定担当者」、「システム安全性」と「システム完全性」などの複合語パターンを抽出し、ノイズ情報Nを生成する。図8に示すように、ノイズ情報Nは、例えばリストである。

40

【0094】

ここで、図8に示す「集積地区」と「各集積地区」、「対象エリア」と「元対象エリア」は、図6に示すパターン情報として登録された接頭語「各」、「元」、や、接尾語「前」、「時」、が一方の単語に付随する複合語のノイズ情報を示す。また、図8に示す「設置担当者」と「設定担当者」、「システム安全性」と「システム完全性」などは、図7に示すパターン情報として登録された意味が近くない一文字違いの熟語の組み合わせ、が含まれている複合語のノイズ情報を示す。

50

【0095】

次に、分析部20'は、単語情報表に、ノイズ情報を利用して、暫定順位として上位X位（Xは任意の自然数）までにおけるパターン情報に該当する単語組み合わせの割合である上位ノイズ率（例えば、上位ノイズ率R）を算出する。そして、分析部20'は所定の誤字脱字抽出条件変更ルールに基づき、誤字脱字語Waを抽出する順位の下限値である下限順位S（Sは任意の自然数）を決定する。この動作は、上述した本発明の第1の実施形態における文書分析装置100の誤字脱字抽出条件変更部24に相当する。

【0096】

例えば、図5のケースでは、暫定順位で上位10位までに誤検出パターン情報Cに該当する単語組み合わせが4～7、10位の計5組存在し、上位ノイズ率Rは0.5（50%）となる。上位ノイズ率Rと下限順位Sの関係式は以下の式（3）などが有効である。

$$S = \gamma / R \quad (3)$$

ここで、 γ は、例えば、図5に示す単語情報表の抽出順位とみなすことができ、文書によらず定数を設定してもよい。また、 γ は、文書Dの文字数や単語数などと単調増加の関係となるように設定するのが有効である。図5のケースで、 $\gamma = 4$ と設定した場合、下限順位Sは8となり、上位8位までが抽出する順位となる。

【0097】

さらに、分析部20'は、誤字度IAが大きい順に並んでいる単語情報表から、ノイズ情報を利用して、パターン情報に該当する単語の組み合わせを除外し、下限順位S位までの組み合わせを、正しいと推定される単語Warおよび誤字脱字語Waの組み合わせとして抽出する。この動作は、上述した本発明の第1の実施形態における文書分析装置100の誤字脱字抽出部25に相当する。

【0098】

図9及び図10は、本発明の第2の実施形態における誤字脱字に関する情報の具体例を示す図である。例えば、図5のケースでは、分析部20'は誤検出パターン情報Cに該当する4～7、10～14、18位の単語組み合わせを除外し、除外後の最終順位の下限順位として算出した8位までを抽出することで、図9のような過不足の少ないリストを得る。

【0099】

一方、文書Dがより誤字脱字に気をかけて記載された文章で「登録データ」、「ハードディスク」、「出データ形式」、「登録情報参照画面」が存在しなかった場合、上位ノイズ率Rは0.8（80%）となり、 $\gamma = 4$ と設定した場合、下限順位Sは5となり、上位5位までが抽出する順位となる。誤検出パターン情報Cに該当する単語組み合わせを除外し、除外後の最終順位の下限順位として算出した5位までを抽出することで、より抽出する情報を限定しながら、図10のような過不足の少ないリストを得る。

【0100】

さらに、分析部20'は、各誤字脱字語Waの文書中の位置や誤字度IA、正しいと推定される単語Warを誤字脱字情報Aとして、分析実施者Kの持つPC端末上に提示する。例えば、図9、図10のようなリストをそのまま提示してもよいし、誤字脱字語Waの文番号に基づき、文書D内における各誤字脱字語Waを着色し明示することで、修正すべき誤字脱字の箇所を分かりやすくし表示してもよい。また、誤字脱字情報Aに基づき、誤字脱字語Waの出現数を文書D全体および、目次の章単位などで集計して表もしくはグラフなどの形式で出力することで、文書Dの品質を表すメトリクスおよび、修正すべき章を判断する情報を提供してもよい。

【0101】

上記のとおり、本発明の第2の実施形態において、文書分析システム300は、各誤字脱字語Waの文書中の位置や誤字度IA、正しいと推定される単語Warを基に誤字脱字に関する情報（例えば、誤字脱字情報A）を抽出し、出力する。したがって、当該文書分析システム300は、文書の内容に依存する誤字脱字を抽出することができる。また、当該文書分析システム300は、少なくとも1つの文書（例えば、文書D）における誤字脱

10

20

30

40

50

字の把握を容易にし、文書の改善を効率化することができる。

【0102】

＜第3の実施形態＞

本発明を実施するための第3の形態における文書分析装置1000の構成について、図面を参照して説明する。なお、本発明の第3の実施形態において、本発明の第1の実施形態及び第2の実施形態と同様の構成については、説明を省略する。

【0103】

図11は、本発明の第3の実施形態における、文書分析装置1000の構成例を示す図である。図11に示すように文書分析装置1000は、誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶する記憶部400と、分析部200とを含む。分析部200は、少なくとも1つの文書からその文書が有する単語及び単語の組み合わせを複数含む単語情報を抽出し、単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、単語情報、誤字脱字指標及びパターン情報に基づいて、誤字脱字に関する情報を生成する。

10

【0104】

なお、第3の実施形態における分析部200と、記憶部400とは、それぞれ、上記の実施形態における分析部20と、ノイズデータベース40とに対応する。

【0105】

上記構成を有する文書分析装置1000は、誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶し、少なくとも1つの文書からその文書が有する単語及び単語の組み合わせを複数含む単語情報を抽出する。そして、文書分析装置1000は、単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、単語情報、誤字脱字指標及びパターン情報に基づいて、誤字脱字に関する情報を生成する。そのため、文書分析装置1000は、文書の内容に依存する誤字脱字を抽出することができる。

20

【0106】

以上、実施形態を用いて本願発明を説明したが、本願発明は、上記実施形態に限定されるものではない。本願発明の構成や詳細には、本願発明の範囲内で当業者が理解しうる様々な変更をすることができる。

【0107】

上記の実施形態の一部又は全部は、以下の付記のようにも記載されうるが、以下には限られない。

30

【0108】

[付記1]

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶する記憶部と、

少なくとも1つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する分析部と、

を含む文書分析装置。

40

【0109】

[付記2]

前記誤字脱字に関する情報は、少なくとも前記文書が有する正しい単語と、前記正しい単語に対応する誤字脱字を含む単語と、前記誤字脱字を含む単語に対する誤字脱字の可能性を示す前記誤字脱字指標との対応関係を含み、

前記分析部は、

前記誤字脱字に関する情報から抽出する前記対応関係の数を算出し、算出した前記対応関係の数を前記誤字脱字に関する情報に含め、前記誤字脱字指標の値の大きい順に、前記対応関係の数だけ前記対応関係を抽出する付記1に記載の文書分析装置。

【0110】

50

〔付記 3〕

前記分析部は、文書解析部と、誤字脱字指標生成部と、ノイズ情報検出部と、誤字脱字抽出条件変更部と、誤字脱字抽出部とをさらに有し、前記記憶部はノイズデータベースをさらに有し、

前記文書解析部は、前記文書が有する少なくとも 1 つの文章に形態素解析を適用することで前記単語情報を抽出し、

前記誤字脱字指標生成部は、前記単語情報から前記誤字脱字指標を算出し、

前記ノイズデータベースは、前記誤字脱字指標で前記誤字脱字を含む単語を検出しようとした場合に誤検出となるパターンを抽出可能にする情報を蓄積し、任意の単語の組み合わせが誤検出パターンに該当するかどうかに関する問い合わせに対し、問い合わせ対象の単語の組み合わせが誤検出パターンと一致するか否かを示す情報を応答し、

前記ノイズ情報検出部は、前記誤字脱字指標に基づき、複数の前記単語の組み合わせに対して、前記誤字脱字指標で誤検出となるパターンの有無を前記ノイズデータベースに問合せ、ノイズとなる前記単語の組み合わせをノイズ情報として検出し、

前記誤字脱字抽出条件変更部は、前記ノイズ情報を利用して所定の誤字・脱字抽出条件変更ルールに基づき、前記誤字脱字を含む単語の抽出条件を変更し、

前記誤字脱字抽出部は、前記文書内の全ての前記単語の組み合わせから、前記ノイズ情報に該当する前記単語の組み合わせを除外し、前記抽出条件に基づき、前記誤字脱字を含む単語を前記誤字脱字に関する情報として抽出する付記 1 又は 2 に記載の文書分析装置。

【0111】

〔付記 4〕

前記誤字脱字指標は、事前に前記文書で使用可能な単語の一覧を登録した使用単語辞書に登録された登録語と、前記文書解析部が抽出した前記単語情報に含まれる複数の単語の各々との文字列の類似性に基づき、前記類似性が近いほど値が高くなる指標である付記 3 に記載の文書分析装置。

【0112】

〔付記 5〕

前記パターン情報は、前記抽出された前記単語情報に含まれる複数の単語の各々と、前記登録語との一致状況において、前記抽出された各単語が類似する文字列があるとした登録語とは別に完全一致する登録語があるパターンを含む付記 4 に記載の文書分析装置。

【0113】

〔付記 6〕

前記誤字脱字抽出部は、前記単語の組み合わせが前記登録語又は前記抽出された前記単語の組み合わせの場合、前記抽出された前記単語を誤字脱字語として抽出することを特徴とする付記 4 又は 5 に記載の文書分析装置。

【0114】

〔付記 7〕

前記誤字脱字抽出部は、誤字脱字抽出条件変更ルールとして、前記誤字脱字指標に基づき、前記単語の組み合わせを前記誤字脱字指標の高い順に並べたリストを作成し、前記リストにおいて上位に並べられた前記単語の組み合わせのうち所定の数の前記単語の組み合わせ、又は、前記誤字脱字指標が一定値以上の前記単語の組み合わせにおいて、前記ノイズ情報に該当した前記単語の組み合わせの割合に対して単調減少の関係にある指標に基づき、前記抽出する前記単語の組み合わせの数を変更し、上位に並べられた前記単語の組み合わせから、前記変更した前記単語の組み合わせ数分の前記単語の組み合わせを新たに抽出し、複数の前記単語の組み合わせのうち出現頻度が低い単語を誤字脱字語として抽出する、ことを特徴とする付記 3 乃至 6 のいずれかに記載の文書分析装置。

【0115】

〔付記 8〕

前記誤字脱字抽出部は、前記誤字脱字抽出条件変更ルールとして、前記誤字脱字指標に基づき、前記単語の組み合わせを前記誤字脱字指標の高い順に並べたリストを作成し、前

10

20

30

40

50

記リストにおいて上位に並べられた所定の数の前記単語の組み合わせ、又は、前記誤字脱字指標が一定値以上の前記単語の組み合わせに対して、前記ノイズ情報に該当した前記単語の組み合わせの割合に対して単調増加の関係にある指標に基づき、前記抽出する前記単語の組み合わせの前記誤字脱字指標の下限値を変更し、上位に並べられた前記単語の組み合わせから、変更した前記下限値までの組み合わせを抽出し、複数の前記単語の組み合わせのうち出現頻度が低い単語を誤字脱字語として抽出する、ことを特徴とする付記 3 乃至 7 のいずれかに記載の文書分析装置。

【0116】

[付記 9]

前記誤字脱字指標は、単語間の文字列の類似性に基づき、前記単語間の文字列の類似性が近いほど、値が高くなる指標であることを特徴とする付記 1 乃至 8 のいずれかに記載の文書分析装置。

10

【0117】

[付記 10]

前記誤字脱字指標は、前記文書における前記単語の出現頻度の偏りに基づき、前記文書における前記出現頻度の偏りが大きいほど、値が高くなる指標である付記 1 乃至 9 のいずれかに記載の文書分析装置。

【0118】

[付記 11]

前記パターン情報は、接辞語の語群を含み、前記単語間の文字列の類似性が近くなる接辞語が一方に付随するパターンを含む付記 1 乃至 10 のいずれかに記載の文書分析装置。

20

【0119】

[付記 12]

前記パターン情報は、意味が類似しない少なくとも一文字違いの熟語の組み合わせを含み、複数の名詞又は未知語を結合した複合語において前記一文字違いの熟語が異なるだけの組み合わせたパターンを含む付記 1 乃至 11 のいずれかに記載の文書分析装置。

【0120】

[付記 13]

前記文書を受け付ける入力部と、前記誤字脱字に関する情報を出力する出力部と、をさらに有する付記 1 乃至 12 のいずれかに記載の文書分析装置。

30

【0121】

[付記 14]

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を含むサーバ装置と、

少なくとも 1 つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する分析部を含む文書分析装置と、を含む文書分析システム。

【0122】

40

[付記 15]

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報を記憶し、

少なくとも 1 つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する文書分析方法。

【0123】

[付記 16]

誤字脱字を含む単語の組み合わせの中で誤字脱字を引き起こす要因となるパターン情報

50

を記憶する処理と、

少なくとも1つの文書から前記文書が有する単語及び前記単語の組み合わせを複数含む単語情報を抽出し、前記単語情報から誤字脱字の可能性を示す誤字脱字指標を算出し、前記単語情報、前記誤字脱字指標及び前記パターン情報に基づいて、誤字脱字に関する情報を生成する処理と、

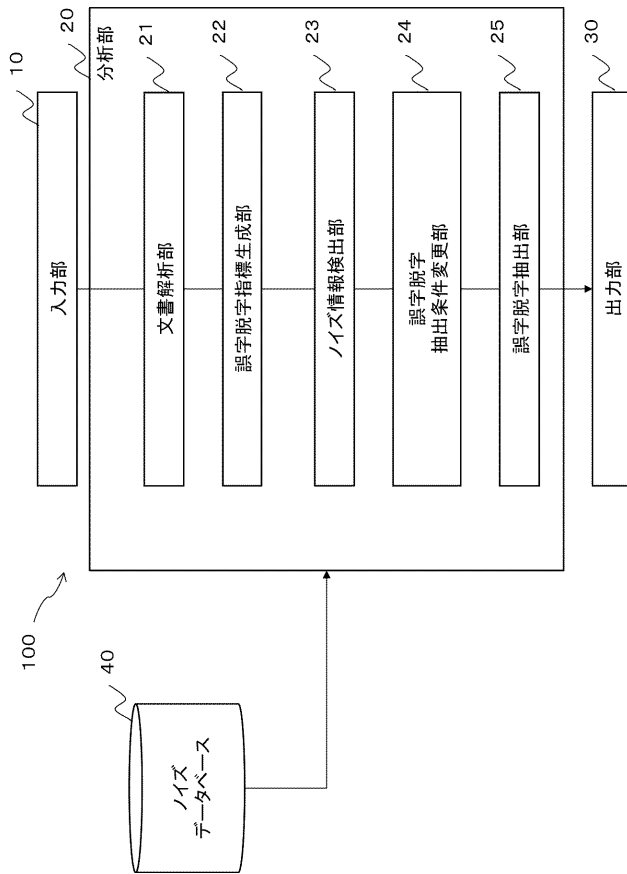
をコンピュータに実行させるプログラム。

【符号の説明】

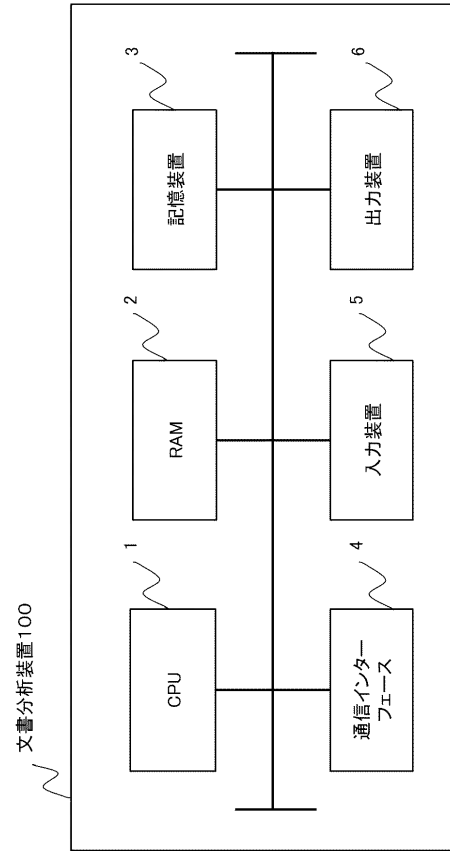
【0124】

1	CPU	
2	RAM	10
3	記憶装置	
4	通信インターフェース	
5	入力装置	
6	出力装置	
10	入力部	
20	分析部	
20'	分析部	
21	文書解析部	
22	誤字脱字指標生成部	
23	ノイズ情報検出部	20
24	誤字脱字抽出条件変更部	
25	誤字脱字抽出部	
30	出力部	
40	ノイズデータベース	
100	文書分析装置	
100'	文書分析装置	
110	サーバ装置	
200	分析部	
300	文書分析システム	
1000	文書分析装置	30

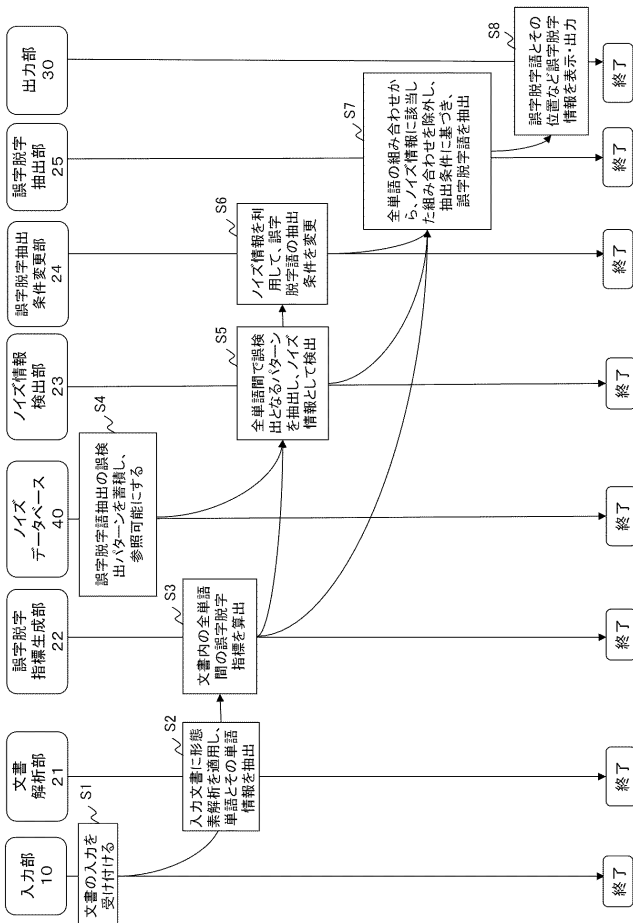
【図 1】



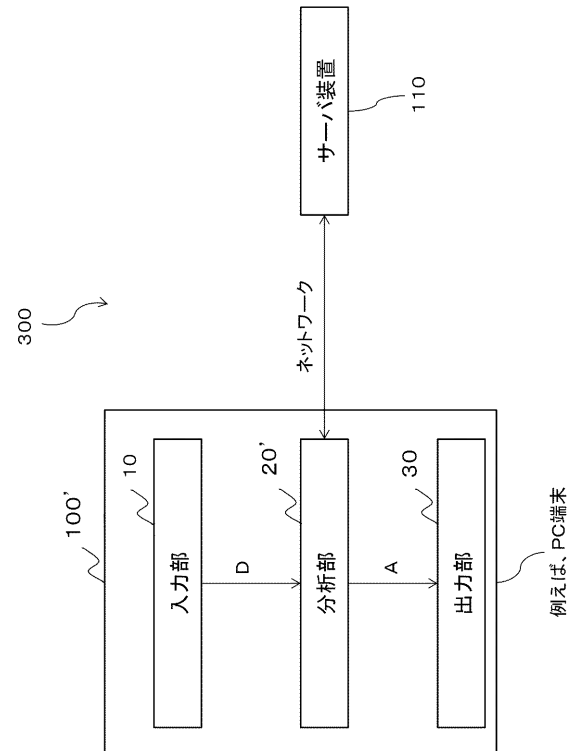
【図 2】



【図 3】



【図 4】



【図 5】

暫定 順位	正しいと推定される単語 Warの候補	誤字脱字語Wa の候補	Warの出現個数 m _{War}	誤字脱字語Wa の候補	Waの出現個数 m _{Wa}	誤字度 IA
1	単語出現頻度	単語発現頻度	46	単語発現頻度	1	0.98
2	登録済データ	登録データ	26	登録データ	1	0.96
3	ハードディスク	ハードディスク	26	ハードディスク	1	0.96
4	集積地区	各集積地区	18	各集積地区	1	0.95
5	入荷金額	入荷金額等	35	入荷金額等	2	0.95
6	データ入力	データ入力時	33	データ入力時	2	0.94
7	対象エリア	元対象エリア	25	元対象エリア	2	0.93
8	出力データ形式	出力データ形式	11	出力データ形式	1	0.92
9	入荷金額	入荷金額	21	入荷金額	2	0.91
10	データ一括選択	データ一括選択前	29	データ一括選択前	3	0.91
11	設置担当者	設定担当者	18	設定担当者	2	0.90
12	入力データ形式	出力データ形式	9	出力データ形式	1	0.90
13	緊急運用準備	緊急費用準備	7	緊急費用準備	1	0.88
14	システム安全性	システム完全性	19	システム完全性	3	0.86
15	集積地区	集積地	15	集積地	3	0.83
16	入力情報参照画面	登録情報参照画面	28	登録情報参照画面	1	0.82
17	検索エリア指定機能	検索エリア登録機能	22	検索エリア登録機能	1	0.81
18	変換可能	変換機能	8	変換機能	2	0.80
19	店舗コード読込処理	店舗コード入力処理	11	店舗コード入力処理	2	0.70
20	.	.	.	.	.	.

【図 6】

接頭語	接尾語
各	前
元	時
再	等
.	.
.	.

【図 7】

一文字違いの熟語の組み合わせ	
設置	設定
運用	費用
安全	完全
入力	出力
.	.
.	.

【図 8】

暫定 順位	正しいと推定される単語 Warの候補	誤字脱字語Wa の誤検出	理由	誤字度 IA
4	集積地区	各集積地区	接頭語	0.95
5	入荷金額	入荷金額等	接尾語	0.95
6	データ入力	データ入力時	接尾語	0.94
7	対象エリア	元対象エリア	接頭語	0.93
10	データ一括選択	データ一括選択前	接尾語	0.91
11	設置担当者	設定担当者	一文字違いの熟語	0.90
12	入力データ形式	出力データ形式	一文字違いの熟語	0.90
13	緊急運用準備	緊急費用準備	一文字違いの熟語	0.88
14	システム安全性	システム完全性	一文字違いの熟語	0.86
18	変換可能	変換機能	一文字違いの熟語	0.80

【図 9】

最終 順位	暫定 順位	正しいと推定される 単語War	Warの出現個数 m _{War}	誤字脱字語Wa	Waの出現 個数m _{Wa}	誤字度 IA
1	1	単語出現頻度	46	単語発現頻度	1	0.98
2	2	登録済データ	26	登録データ	1	0.96
3	3	ハードディスク	26	ハードディスク	1	0.96
4	8	出力データ形式	11	出力データ形式	1	0.92
5	9	入荷金額	21	入荷金額	2	0.91
6	15	集積地区	15	集積地	3	0.83
7	16	入力情報参照画面	28	登録情報参照画面	1	0.82
8	17	検索エリア指定機能	22	検索エリア登録機能	1	0.81

【図 1 0】

最終 順位	暫定 順位	正しいと推定される 単語War	Warの出現個数 m _{War}	誤字脱字語Wa	Waの出現 個数m _{Wa}	誤字度 IA
1	1	単語出現頻度	46	単語発現頻度	1	0.98
2	9	入荷金額	21	入荷額	2	0.91
3	15	集積地区	15	集積地	3	0.83
4	17	検索エリア指定機能	22	検索エリア登録機能	1	0.81
5	19	店舗コード読込処理	11	店舗コード入力処理	2	0.70

【図 1 1】

