(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第5362353号
(P5362353)

(45)発行日　平成25年12月11日(2013.12.11)　(24)登録日　平成25年9月13日(2013.9.13)

(51)Int.Cl.　　F I
G06F 17/21 (2006.01)　　G06F 17/21 550J
G06F 17/30 (2006.01)　　G06F 17/21 570N
G06F 17/30 170A
G06F 17/30 330C

請求項の数 6　(全 12 頁)

(21)出願番号　特願2008-520339 (P2008-520339)
(86)(22)出願日　平成18年6月30日 (2006.6.30)
(65)公表番号　特表2009-500754 (P2009-500754A)
(43)公表日　平成21年1月8日 (2009.1.8)
(86)国際出願番号　PCT/US2006/026012
(87)国際公開番号　WO2007/008492
(87)国際公開日　平成19年1月18日 (2007.1.18)
審査請求日　平成21年6月19日 (2009.6.19)
(31)優先権主張番号　11/177,136
(32)優先日　平成17年7月8日 (2005.7.8)
(33)優先権主張国　米国 (US)

(73)特許権者　500046438
マイクロソフト　コーポレーション
アメリカ合衆国　ワシントン州　９８０５２－６３９９　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100107766
弁理士　伊東　忠重
(74)代理人　100070150
弁理士　伊東　忠彦
(74)代理人　100091214
弁理士　大貫　進介

最終頁に続く

(54)【発明の名称】文書中のコロケーション誤りを処理すること

(57)【特許請求の範囲】
【請求項１】
文書中のコロケーション誤りを訂正するためのコンピュータにより実施される方法であって、
前記文書中のテキストの文にアクセスするステップと、
前記テキストの前記文において１または複数の検査語を検出するステップであって、前記検査語は、コロケーション誤りを生じさせる可能性がある語であって、特定のコロケーション誤りタイプを示すこと、
検出された検査語に基づいて前記コロケーション誤りタイプを識別するステップと、
前記検出された検査語により示された特定のコロケーション誤りタイプに基づいてクエリを生成するステップであって、メモリ内の複数の異なる組のクエリ生成ルールから一組のルールを選択し、前記検査語が含まれた特定の組のクエリを生成するステップと、
前記特定の組のクエリを、検索モジュールにサブミットして検索結果を得るステップであって、前記検索結果は、ウェブ検索エンジンを用いて抽出された文書のテキストのサマリーを含むこと、
前記検索結果のテキストと前記特定の組のクエリのテキストとを比較して、１または複数のクエリのテキストが一致した場合には、コロケーション誤りが存在しないと判定し、全てのクエリのテキストが一致しなかった場合には、コロケーション誤りがあると判定するステップと、
コロケーション誤りがあると判定した場合に、前記コロケーション誤りタイプに基づい

て、前記検査語に対応する語をプレースホルダーに置換したクエリテンプレートを生成するステップと、
前記クエリテンプレートを、前記検索モジュールにサブミットして文字列を得るステップと、
前記クエリテンプレートと一致する文字列であって、前記プレースホルダーが任意の語と一致する文字列を識別するステップと、
前記識別された文字列を表示して、前記検出されたコロケーション誤りに対する訂正の選択肢を提供するステップと
を備えたことを特徴とする方法。
【請求項２】
前記検査語を検出するステップは、前記文を構文解析して、前記文に含まれる品詞を識別するステップをさらに含み、
前記クエリを生成するステップは、識別された前記品詞に基づくことを特徴とする請求項１に記載の方法。
【請求項３】
前記識別された文字列を、前記クエリテンプレートに対応する重みに基づいてランク付けするステップをさらに含むことを特徴とする請求項１に記載の方法。
【請求項４】
前記コロケーション誤りタイプは、動詞－名詞、前置詞－名詞、形容詞－名詞、および動詞－副詞のうちの少なくとも１つを含むことを特徴とする請求項１に記載の方法。
【請求項５】
前記特定の組のクエリは、
前記テキストの文を含む文クエリと、
前記テキストの文のチャンクを含むチャンククエリと、
前記テキストの文の主要語ペアを含む語クエリと
を含むことを特徴とする請求項１に記載の方法。
【請求項６】
前記検索モジュールは、ウェブベースの検索エンジンであることを特徴とする請求項１に記載の方法。
【発明の詳細な説明】
【背景技術】
【０００１】
以下の説明は、単に、一般的な背景情報として提供され、主張される主題の範囲を確定する助けとして使用されることを意図するものではない。
【０００２】
成長の一途をたどるグローバル経済、およびインターネットの急速な発展とともに、世界中の人々が、その人々の母語ではない言語で書くことにますます慣れ親しんでいる。残念ながら、大きく異なる文化および文体を有する、いくつかの社会に関して、いくつかの母語ではない言語で書く能力は、常に存在する障壁である。母語ではない言語（例えば、英語）で書く際、言語使用の誤りが、ノンネイティブスピーカ（例えば、中国語、日本語、韓国語その他の英語ではない他の言語を話す人々）によって頻繁に犯される。この種の誤りには、文法上の誤りと、動詞－目的語、形容詞－名詞、副詞－動詞等のコロケーションの不適切な使用の両方が含まれる。
【０００３】
多くの人々は、適切な文法を使用して母語ではない言語で書く能力を有するが、それらの人々は依然として、２つの語の間のコロケーションの誤りに苦労する可能性がある。さらに他の人々は、文法と、２つの語の間のコロケーションなどの他の誤りとの両方に苦労する。スペルチェックプログラムおよび文法チェックプログラムが文法上の誤りを訂正するのに役立つとはいえ、２つの語の間のコロケーションの誤りの検出および／または訂正は、特にこれらの誤りがその他の点で文法的に正しい可能性があるので、困難である可能

性がある。したがって、文法チェッカは通常、語の間のコロケーションと関係する誤りを検出するのに全くと言っていいほど助けとならない。以下の説明において、母語ではない言語の例として英語が使用されるが、以上の問題は、他の言語境界にも存在する。

【０００４】

例えば、その他の点で文法的に正しい場合でも、文をネイティブらしい英語でなくするコロケーション誤りを含む、以下の文を考える。すなわち、

１．Ｏｐｅｎ ｔｈｅ ｌｉｇｈｔ．
２．Ｅｖｅｒｙｂｏｄｙ ｈａｔｅｓ ｔｈｅ ｃｒｏｗｄｅｄ ｔｒａｆｆｉｃ ｏｎ ｗｅｅｋｅｎｄｓ．
３．Ｔｈｉｓ ｉｓ ａ ｃｈｅｃｋ ｏｆ ＵＳ＄５００．
４．Ｉ ｃｏｎｇｒａｔｕｒａｔｅ ｙｏｕ ｆｏｒ ｙｏｕｒ ｓｕｃｃｅｓｓ．

以上の文のネイティブらしい英語のバージョンは、以下のようでなければならない。すなわち、

１．Ｔｕｒｎ ｏｎ ｔｈｅ ｌｉｇｈｔ．
２．Ｅｖｅｒｙｂｏｄｙ ｈａｔｅｓ ｔｈｅ ｈｅａｖｙ ｔｒａｆｆｉｃ ｏｎ ｗｅｅｋｅｎｄｓ．
３．Ｔｈｉｓ ｉｓ ａ ｃｈｅｃｋ ｆｏｒ ＵＳ＄５００．
４．Ｉ ｃｏｎｇｒａｔｕｒａｔｅ ｙｏｕ ｏｎ ｙｏｕｒ ｓｕｃｃｅｓｓ．

である。

【０００５】

英語を母語としない話者が直面する障壁の例として、中国人ユーザの窮状を検討されたい。文化、背景、および考え方の習慣により、中国人はしばしば、文法的ではあるが自然ではない英文を作る。例えば、中国人は、中国語の主語を英語の主語に直接に翻訳し、目的語および動詞に関しても、同じことを行う傾向がある。英語で書く際、中国人は、しばしば、動詞と前置詞の間、形容詞と名詞の間、動詞と名詞の間等のコロケーションを決定するのに困難を覚える。さらに、ビジネス分野等の特定の分野では、特別なライティングスキル及びライティングスタイルが必要とされる。

【発明の開示】

【発明が解決しようとする課題】

【０００６】

一般的な辞書は、リーディング（一種の復号プロセス）の目的でノンネイティブスピーカによって主に使用されるが、これらの辞書は、ライティング（一種の符号化プロセス）のための十分な支援を提供しない。これらの辞書は、単一の語の説明を提供するだけであり、通常、関係のある句およびコロケーションを説明する十分な情報を提供しない。さらに、この種の情報を、その情報のいくらかが辞書において提供されている場合でさえ、辞書から簡単に得る方法が存在しない。他方、現在広く使用されている文法チェックツールは、犯しやすい文法上の誤りを検出する、いくらかの限られた能力を有するが、コロケーション誤りを検出することができない。

【課題を解決するための手段】

【０００７】

この概要は、発明を実施するための最良の形態において後述されるいくつかの概念を簡略化された形で導入するために提供される。この概要は、主張される主題の重要な特徴または不可欠な特徴を特定することを意図しておらず、また、主張される主題の範囲を確定する助けとして使用されることも意図していない。

【０００８】

文がアクセスされ、少なくとも１つのクエリが、前記文に基づいて生成される。前記少なくとも１つのクエリは、例えばウェブ検索エンジンを使用して、ある文書コレクション内のテキストと比較されることが可能である。前記文中のコロケーション誤りが、前記少なくとも１つのクエリと前記文書コレクション内のテキストとの比較に基づき、検出され、かつ／または訂正されることが可能である。

【発明を実施するための最良の形態】
【０００９】
図１は、本発明を実施することができる適切なコンピューティングシステム環境１００の例を示している。コンピューティングシステム環境１００は、適切なコンピューティング環境の一例に過ぎず、本発明の用法または機能の範囲について何ら限定を示唆することを意図していない。また、コンピューティング環境１００が、例示的な動作環境１００に示されるコンポーネントのいずれの１つ又は組合せに関連する依存関係または要件も有すると解釈してはならない。
【００１０】
本発明は、他の多数の汎用または専用のコンピューティングシステム環境またはコンピューティングシステム構成で動作する。本発明で使用するのに適する可能性がある周知のコンピューティングシステム、コンピューティング環境、および／またはコンピューティング構成の例には、これらに限定されないが、パーソナルコンピュータ、サーバコンピュータ、ハンドヘルドデバイスまたはラップトップデバイス、マルチプロセッサシステム、マイクロプロセッサベースのシステム、セットトップボックス、プログラマブル家庭用電子機器、ネットワークＰＣ、ミニコンピュータ、メインフレームコンピュータ、電話システム、以上のシステム又はデバイスのいずれかを含む分散コンピューティング環境等が含まれる。
【００１１】
本発明は、コンピュータによって実行される、プログラムモジュールなどのコンピュータ実行可能命令の一般的な文脈において説明することができる。一般に、プログラムモジュールには、特定のタスクを実行する又は特定の抽象データ型を実装する、ルーチン、プログラム、オブジェクト、コンポーネント、データ構造等が含まれる。また本発明は、通信ネットワークを介してリンクされたリモート処理装置によってタスクが実行される、分散コンピューティング環境で実施してもよい。分散コンピューティング環境では、プログラムモジュールは、メモリ記憶装置を含むローカルコンピュータ記憶媒体とリモートコンピュータ記憶媒体の両方の中に配置されることが可能である。プログラムおよびモジュールによって実行されるタスクについては、後段で、図の助けを借りて説明する。当業者は、任意の形のコンピュータ読み取り可能な媒体上に書き込むことが可能なプロセッサ実行可能命令として、説明および図を実施することができる。
【００１２】
図１を参照すると、本発明を実施するための例示的システムは、コンピュータ１１０の形で汎用コンピューティングデバイスを含む。コンピュータ１１０のコンポーネントには、これらに限定されないが、処理装置１２０、システムメモリ１３０、及びシステムメモリを含む様々なシステムコンポーネントを処理装置１２０に結合するシステムバス１２１が含まれる。システムバス１２１は、様々なバスアーキテクチャのいずれかを使用する、メモリバスまたはメモリコントローラ、周辺バス、およびローカルバスを含め、いくつかのタイプのバス構造のいずれであってもよい。例として、限定としてではなく、そのようなアーキテクチャには、ＩＳＡ（Ｉｎｄｕｓｔｒｙ Ｓｔａｎｄａｒｄ Ａｒｃｈｉｔｅｃｔｕｒｅ）バス、ＭＣＡ（Ｍｉｃｒｏ Ｃｈａｎｎｅｌ Ａｒｃｈｉｔｅｃｔｕｒｅ）バス、ＥＩＳＡ（Ｅｎｈａｎｃｅｄ ＩＳＡ）バス、ＶＥＳＡ（Ｖｉｄｅｏ Ｅｌｅｃｔｒｏｎｉｃｓ Ｓｔａｎｄａｒｄｓ Ａｓｓｏｃｉａｔｉｏｎ）ローカルバス、およびメザニン（Ｍｅｚｚａｎｉｎｅ）バスとしても知られるＰＣＩ（Ｐｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ）バスが含まれる。
【００１３】
コンピュータ１１０は通常、様々なコンピュータ読み取り可能な媒体を備える。コンピュータ読み取り可能な媒体は、コンピュータ１１０がアクセスすることができる任意の利用可能な媒体であることが可能であり、揮発性媒体と不揮発性媒体、リムーバブルな媒体とノンリムーバブルな媒体がともに含まれる。例として、限定としてではなく、コンピュータ読み取り可能な媒体は、コンピュータ記憶媒体および通信媒体を含むことが可能であ

る。コンピュータ記憶媒体には、コンピュータ読み取り可能な命令、データ構造、プログラムモジュールその他のデータなどの情報の格納のために任意の方法または技術で実装された、揮発性および不揮発性の、リムーバブルおよびノンリムーバブルな媒体が含まれる。コンピュータ記憶媒体には、これらに限定されないが、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリその他のメモリ技術、ＣＤ－ＲＯＭ、ＤＶＤ（デジタルバーサタイルディスク）その他の光ディスクストレージ、磁気カセット、磁気テープ、磁気ディスクストレージその他の磁気記憶装置、または所望の情報を格納するのに使用することができ、コンピュータ１１０がアクセスすることができる他の任意の媒体が含まれる。通信媒体は通常、搬送波等の変調されたデータ信号の中に、または他のトランスポート機構で、コンピュータ読み取り可能な命令、データ構造、プログラムモジュールその他のデータを実体化し、通信媒体には、あらゆる情報配信媒体（ｉｎｆｏｒｍａｔｉｏｎ　ｄｅｌｉｖｅｒｙ　ｍｅｄｉａ）が含まれる。「変調されたデータ信号」という用語は、信号内に情報を符号化するように特性の１つ又は複数が設定または変更されている信号を意味する。例として、限定としてではなく、通信媒体には、有線ネットワークまたは直接有線接続等の有線媒体、および音響媒体、ＲＦ媒体、赤外線媒体その他の無線媒体等の無線媒体が含まれる。また、以上の媒体のいずれかの媒体の組合せも、コンピュータ読み取り可能な媒体の範囲内に含められなければならない。

【００１４】

システムメモリ１３０は、ＲＯＭ（読み取り専用メモリ）１３１やＲＡＭ（ランダムアクセスメモリ）１３２などの揮発性メモリおよび／または不揮発性メモリの形で、コンピュータ記憶媒体を備える。起動中等に、コンピュータ１１０内部の要素間で情報を転送するのを助ける基本ルーチンを含むＢＩＯＳ（基本入出力システム）１３３が、通常ＲＯＭ１３１の中に格納される。ＲＡＭ１３２は通常、処理装置１２０が、即時にアクセスすることができ、かつ／または現在、処理しているデータおよび／またはプログラムモジュールを含む。限定ではなく例として、図１は、オペレーティングシステム１３４、アプリケーションプログラム１３５その他のプログラムモジュール１３６、およびプログラムデータ１３７を示している。

【００１５】

またコンピュータ１１０は、他のリムーバブル／ノンリムーバブル、揮発性／不揮発性のコンピュータ記憶媒体も備えることが可能である。単に例として、図１は、ノンリムーバブルな不揮発性の磁気媒体に対して読み取り又は書き込みを行うハードディスクドライブ１４１、リムーバブルな不揮発性の磁気ディスク１５２に対して読み取り又は書き込みを行う磁気ディスクドライブ１５１、およびＣＤ－ＲＯＭその他の光媒体等のリムーバブルな不揮発性の光ディスク１５６に対して読み取り又は書き込みを行う光ディスクドライブ１５５を示している。例示的動作環境において使用することができる、他のリムーバブル／ノンリムーバブル、揮発性／不揮発性のコンピュータ記憶媒体には、これらに限定されないたが、磁気テープカセット、フラッシュメモリカード、ＤＶＤ、デジタルビデオテープ、ソリッドステートＲＡＭ、ソリッドステートＲＯＭなどが含まれる。ハードディスクドライブ１４１は通常、インタフェース１４０のようなノンリムーバブルなメモリインタフェースを介してシステムバス１２１に接続され、磁気ディスクドライブ１５１および光ディスクドライブ１５５は通常、インタフェース１５０のようなリムーバブルなメモリインタフェースでシステムバス１２１に接続される。

【００１６】

前述し図１に示したドライブ及び関連するコンピュータ記憶媒体は、コンピュータ読み取り可能な命令、データ構造、プログラムモジュールその他のデータのストレージをコンピュータ１１０に提供する。図１では例えば、ハードディスクドライブ１４１が、オペレーティングシステム１４４、アプリケーションプログラム１４５、他のプログラムモジュール１４６、およびプログラムデータ１４７を格納しているものとして示されている。これらのコンポーネントは、オペレーティングシステム１３４、アプリケーションプログラム１３５、他のプログラムモジュール１３６、およびプログラムデータ１３７と同一であ

ることも異なることも可能であることに留意されたい。オペレーティングシステム１４４、アプリケーションプログラム１４５、他のプログラムモジュール１４６、およびプログラムデータ１４７には、少なくともそれらが異なるコピーであることを示すために、異なる符号を与えている。

【００１７】

ユーザは、キーボード１６２、マイクロホン１６３、および、マウス、トラックボール又はタッチパッドなどのポインティングデバイス１６１などの入力デバイスを介して、コマンドおよび情報をコンピュータ１１０に入力することができる。他の入力デバイス（図示せず）としては、ジョイスティック、ゲームパッド、サテライトディッシュ、スキャナなどを含めることが可能である。これら及びその他の入力デバイスはしばしば、システムバスに結合されたユーザ入力インタフェース１６０を介して処理装置１２０に接続されるが、パラレルポート、ゲームポート、またはＵＳＢ（ユニバーサルシリアルバス）などの、他のインタフェースおよびバス構造で接続してもよい。また、モニタ１９１その他のタイプのディスプレイデバイスも、ビデオインタフェース１９０のようなインタフェースを介してシステムバス１２１に接続される。モニタに加え、コンピュータは、出力周辺インタフェース１９０を介して接続することができる、スピーカ１９７やプリンタ１９６などの、他の周辺出力デバイスも備えることが可能である。

【００１８】

コンピュータ１１０は、リモートコンピュータ１８０のような１つ又は複数のリモートコンピュータに対する論理接続を使用するネットワーク化された環境で動作することができる。リモートコンピュータ１８０は、パーソナルコンピュータ、ハンドヘルドデバイス、サーバ、ルータ、ネットワークＰＣ、ピアデバイスその他の一般的なネットワークノードであることが可能であり、通常、コンピュータ１１０に関連して前述した要素の多く又はすべてを備える。図１に示した論理接続には、ＬＡＮ（ローカルエリアネットワーク）１７１およびＷＡＮ（ワイドエリアネットワーク）１７３が含まれるが、他のネットワークを含むことも可能である。そのようなネットワーキング環境は、オフィス、企業全体のコンピュータ網、イントラネット及びインターネットで一般的である。

【００１９】

ＬＡＮネットワーキング環境で使用される場合、コンピュータ１１０は、ネットワークインタフェース又はアダプタ１７０を介してＬＡＮ１７１に接続される。ＷＡＮネットワーキング環境で使用される場合、コンピュータ１１０は通常、インターネットなどのＷＡＮ１７３を介して通信を確立するためのモデム１７２または他の手段を備える。内部にあることも外部にあることも可能なモデム１７２は、ユーザ入力インタフェース１６０その他の適切な機構を介してシステムバス１２１に接続することができる。ネットワーク化された環境では、コンピュータ１１０に関連して示したプログラムモジュール又はプログラムモジュールの部分は、リモートメモリ記憶装置の中に格納することができる。限定ではなく例として、図１は、リモートアプリケーションプログラム１８５が、リモートコンピュータ１８０上に存在するものとして示す。図示したネットワーク接続は、例示的であり、コンピュータ間で通信リンクを確立する他の手段も使用できることを理解されたい。

【００２０】

図２は、テキスト内のコロケーション誤り（ｃｏｌｌｏｃａｔｉｏｎ　ｅｒｒｏｒ）を検出して訂正するためのシステム２００の流れ図である。多くのタイプのコロケーション誤りが存在する。システム２００の一態様では、４つのタイプのコロケーション誤りが検出される。コロケーション誤りのタイプには、以下が含まれる。すなわち、

１．動詞－名詞（ＶＮ、例えば、＊ｌｅａｒｎ／ａｃｑｕｉｒｅ　ｋｎｏｗｌｅｄｇｅ）

２．前置詞－名詞（ＰＮ、例えば、＊ｏｎ／ｉｎ　ｔｈｅ　ｍｏｒｎｉｎｇ）

３．形容詞－名詞（ＡＮ、例えば、＊ｓｏｃｉａｌ／ｓｏｃｉａｌｉｓｔ　ｃｏｕｎｔｒｙ）、および

４．動詞－副詞（ＶＡ、例えば、ｓｉｔｕａｔｉｏｎｓ　ｃｈａｎｇｅ　＊ｌａｒｇｅｌｙ／ｇｒｅａｔｌｙ）

前処理モジュール２０２が、テキストを処理して、テキストの品詞タグ付け（ｔａｇｇｉｎｇ）および構文解析をもたらす。多くの異なるタイプのパーサを、テキストを処理するのに使用することが可能である。以下は、例示的な文である。

Ｉ ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ ｔｈｉｓ ｐｅｒｓｏｎ ｆｏｒ ｙｅａｒｓ．

前処理モジュール２０２が、この文にタグ付けを行い、この文を以下のとおりに切り分ける。

［ＮＰ Ｉ／ＰＲＰ］［ＶＰ ｈａｖｅ／ＶＢＰ ｒｅｃｏｇｎｉｚｅｄ／ＶＢＮ］［ＮＰ ｔｈｉｓ／ＤＴ ｐｅｒｓｏｎ／ＮＮ］［ＰＰ ｆｏｒ／ＩＮ］［ＮＰ ｙｅａｒｓ． ＜／ｓ＞／ＮＮＳ］

この処理済みのテキストを使用して、クエリ生成モジュール２０４が、クエリを構築する。一例では、４つのクエリセットが、先に特定されたコロケーション誤りタイプの各タイプに関して生成される。例えば、コロケーション誤りタイプは、動詞－名詞、前置詞－名詞、形容詞－名詞、および動詞－副詞であることが可能である。生成されるクエリは、文のフルテキストと共に、補助が削除された文の短縮された部分も含むことが可能である。前掲の文に関する例示的な短縮されたクエリには、「ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ ｔｈｉｓ ｐｅｒｓｏｎ」、「ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ」、「ｔｈｉｓ ｐｅｒｓｏｎ」、および「ｒｅｃｏｇｎｉｚｅｄ ｐｅｒｓｏｎ」を含めることが可能である。

【００２１】

クエリは、検索モジュール２０６にサブミットされる。一実施形態では、検索モジュールは、ＭＳＮ（登録商標） Ｓｅａｒｃｈ（ｓｅａｒｃｈ．ｍｓｎ．ｃｏｍ）、Ｇｏｏｇｌｅ（登録商標）（ｗｗｗ．ｇｏｏｇｌｅ．ｃｏｍ）、および／またはＹａｈｏｏ！（登録商標）（ｗｗｗ．ｙａｈｏｏ．ｃｏｍ）などのウェブベースの検索エンジンであることが可能である。ウェブは、膨大な量のテキストを含むので、コロケーション誤りを検出する安価なリソースであることが可能である。誤り検出モジュール２０８が、クエリ生成モジュール２０４によって生成されたクエリを、検索モジュール２０６によって獲得された結果と比較する。誤り訂正モジュール２１０が、誤り検出モジュール２０８によって識別された誤りに関する候補訂正を提供する。

【００２２】

図３は、図２に示されるシステム２００において実施することが可能な方法２２０の流れ図である。ステップ２２２で、文にアクセスする。文には、ワードプロセッサ、例えば、ワシントン州リッチモンドのマイクロソフトコーポレーションから入手可能なＭｉｃｒｏｓｏｆｔ Ｗｏｒｄ（登録商標）に入力されているテキストが含まれてもよい。ステップ２２４で、その文を構文解析してチャンク（ｃｈｕｎｋ）にし、文中の品詞が識別される。次に、ステップ２２６で、その構文解析に基づいてクエリを生成する。ステップ２２８で、クエリを、ＭＳＮ（登録商標） Ｓｅａｒｃｈ、Ｇｏｏｇｌｅ（登録商標）および／またはＹａｈｏｏ！（登録商標）等の検索エンジンにサブミットする。ステップ２３０で、文中のコロケーション誤りが、それらのクエリと検索エンジンからの結果とを比較することにより検出される。誤りを検出した後、ステップ２３２で、コロケーション誤りの代替物のランク付けされた候補をユーザに提示する。

【００２３】

図４は、図２のクエリ生成モジュール２０４のブロック図である。クエリ生成モジュール２０４は、構文解析済みの文２４０、例えば前処理モジュール２０２から受け取る構文解析済みの文を受け入れる。構文解析済みの文２４０に基づき、クエリ生成モジュール２０４は、文クエリ２４２、チャンククエリ２４４、および語クエリ２４６を生成する。先に特定された可能なコロケーション誤りのタイプを所与として、検査語（ｃｈｅｃｋｉｎｇ ｗｏｒｄ）（すなわち、コロケーション誤りを生じさせる可能性がある語）を、以下のとおり検出する。すなわち、タイプＶＮにおける動詞、タイプＰＮにおける前置詞、タイプＡＮにおける形容詞、およびタイプＶＡにおける副詞である。タイプに応じて、クエリ生成モジュール２０４は、異なるクエリセットを以下のとおり生成する。すなわち、

１．文クエリ２４２：Ｓ－Ｑｕｅｒｙと呼ばれる、元の文と（各タイプに関して予め定められたされた補助（ａｕｘｉｌｉａｒｙ）を削除することにより）短縮された文
２．チャンククエリ２４４：Ｃ－Ｑｕｅｒｙと呼ばれる、文中の対応するチャンクペア、および
３．語クエリ２４６：Ｗ－Ｑｕｅｒｙと呼ばれる、文中の対応する主要語（ｈｅａｄｗｏｒｄ）ペア
である。

【００２４】

タイプＶＮ検出のための、文「Ｉ ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ ｔｈｉｓ ｐｅｒｓｏｎ ｆｏｒ ｙｅａｒｓ」に対する例示的クエリが以下に提示される。～は、２つの隣接する語が、互いに隣接しているか又は１語離れていることが可能であることを意味する。

Ｓ－Ｑｕｅｒｙ：［“Ｉ ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ ｔｈｉｓ ｐｅｒｓｏｎ ｆｏｒ ｙｅａｒｓ”］
Ｓ－Ｑｕｅｒｙ：［“ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ ｔｈｉｓ ｐｅｒｓｏｎ”］
Ｃ－Ｑｕｅｒｙ：［“ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ”～“ｔｈｉｓ ｐｅｒｓｏｎ”］
Ｗ－Ｑｕｅｒｙ：［“ｒｅｃｏｇｎｉｚｅｄ”～“ｐｅｒｓｏｎ”］

【００２５】

各タイプのクエリを生成するための例示的規則は、以下のとおりである。すなわち、
ＶＮ：複数のＳ－Ｑｕｅｒｙ、１つのＣ－Ｑｕｅｒｙ Ｖ～Ｎ、および１つのＷ－Ｑｕｅｒｙ $Ｖ_{h}$～$Ｎ_{h}$（$Ｎ_{h}$は、対応する名詞チャンクの主要語を表す。）、
ＰＮ：前置詞を含む、ＰＮの１つのＣ－Ｑｕｅｒｙ、
ＡＮ：ＡＮペアを含む、ＡＮの１つのＣ－Ｑｕｅｒｙ、および
ＶＡ：ＶＡペアを含む複数のＣ－Ｑｕｅｒｙ、およびＶＡ主要語を含む複数のＷ－Ｑｕｅｒｙ
である。

【００２６】

図５は、ある文中の誤りを検出する方法２５０の流れ図である。クエリ生成モジュール２０４によって生成されたクエリが、ステップ２５１で、検索モジュール２０６にサブミットされる。検索モジュール２０６によって獲得された検索結果が、それらのクエリと比較される。一例では、それらの結果には、ウェブ検索エンジンを使用して取得された文書に関するテキストの要約が含まれる。ステップ２５２で、クエリ生成モジュール２０４からの複数のＳクエリ２４２が、検索モジュールからの結果と比較される。次に、ステップ２５４で、Ｓクエリ２４２の１つ又は複数が、検索モジュール結果と一致するか否かについての判定が行われる。Ｓクエリの１つ又は複数が、検索モジュール結果と一致した場合、ステップ２５６で、コロケーション誤りは全く存在しないものと判定される。

【００２７】

しかし一致が存在しない場合、方法２５０はステップ２５８に進み、複数のＣクエリ２４４が、検索モジュール結果と比較される。ステップ２６０で、Ｃクエリの１つ又は複数が、検索モジュール結果とよく一致するか否か、およびその比較に関するスコアが閾値より大きいか否かが判定される。一例では、このスコアは、Ｃクエリのチャンクが、検索結果の中で出現する回数を、Ｃクエリの中の語がそれらの検索結果の中で同時に出現する回数で割ることによって計算される。そのスコアが、閾値より大きい場合、ステップ２５６で、コロケーション誤りは全く存在しないものと判定される。

【００２８】

そのスコアが、閾値未満である場合、方法２５０はステップ２６２に進み、複数のＷ－Ｑｕｅｒｙが、検索エンジンデータと比較される。ステップ２６４が、それらのＷクエリと検索エンジンデータの間に、よい一致が存在するか否か、および、その比較に関するスコアが、閾値より大きいか否かを判定する。スコアが閾値より大きい場合、ステップ２５

６で、コロケーション誤りは全く存在しないものと判定される。この比較に関するスコアは、Ｃ－Ｑｕｅｒｙ比較スコアと同様であることが可能である。このため、Ｗ－Ｑｕｅｒｙ比較スコアは、Ｗ－Ｑｕｅｒｙが検索結果の中で出現する回数を、Ｗ－Ｑｕｅｒｙの中の語ペアが同時に出現する合計回数で割ることによって計算することが可能である。そのスコアが閾値未満である場合、方法２５０はステップ２６６に進み、ユーザが可能なコロケーション誤りについての通知を受ける。

【００２９】

図６は、可能な訂正済みのコロケーションをユーザに提示するための方法２７０の流れ図である。ステップ２７２で、クエリテンプレートが生成される。クエリテンプレートは、誤りとして識別された語に基づいて生成される（すなわち、前述した検査語は、図５の方法２５０により判定したコロケーション誤りを含む。）。クエリテンプレートは、コロケーション誤りを生じさせる検査語が「＋」で置き換えられた後の入力文から導出される。前出の文において、「ｒｅｃｏｇｎｉｚｅｄ」が、検査語として識別されており、このため、クエリテンプレートはこの語に基づいて展開される。例えば、ＶＮ検出のための、文「Ｉ ｈａｖｅ ｒｅｃｏｇｎｉｚｅｄ ｔｈｉｓ ｐｅｒｓｏｎ ｆｏｒ ｙｅａｒｓ」のクエリテンプレートは、以下のとおりである。＋は、任意の語を表す。

Ｓ－ＱＴ：［“Ｉ ｈａｖｅ＋ｔｈｉｓ ｐｅｒｓｏｎ ｆｏｒ ｙｅａｒｓ”］
Ｓ－ＱＴ：［“Ｉ ｈａｖｅ＋ｔｈｉｓ ｐｅｒｓｏｎ”］
Ｓ－ＱＴ：［“ｈａｖｅ＋ｔｈｉｓ ｐｅｒｓｏｎ ｆｏｒ ｙｅａｒｓ”］
Ｓ－ＱＴ：［“Ｉ ｈａｖｅ＋ｔｈｉｓ ｐｅｒｓｏｎ”］
Ｃ－ＱＴ：［“＋ｔｈｉｓ ｐｅｒｓｏｎ ｆｏｒ ｙｅａｒｓ”］
Ｃ－ＱＴ：［“＋ｔｈｉｓ ｐｅｒｓｏｎ”］

【００３０】

クエリテンプレートを生成するための例示的な規則は、以下のとおりであることが可能である。そなわち、

ＶＮ：Ｓ－ＱＴ、Ｃ－ＱＴ（動詞が＋で置き換えられている。）、
ＰＮ：Ｓ－ＱＴ、Ｃ－ＱＴ（前置詞が＋で置き換えられている。）、
ＡＮ：Ｓ－ＱＴ、Ｃ－ＱＴ（形容詞が＋で置き換えられている。）、および
ＶＡ：Ｓ－ＱＴ、Ｃ－ＱＴ（副詞が＋で置き換えられている。）
である。

【００３１】

ステップ２７４で、クエリテンプレートが、検索モジュールに、本明細書では検索エンジンにサブミットされる。ステップ２７６で、検索エンジン結果からの文字列が取得される。それらの文字列は、取り巻く文脈の語を有するテキストの要約を含むことが可能である。クエリテンプレートと一致する文字列であって、＋の位置は、任意の１つの語であることが可能な文字列が、文字列候補として識別される。そのコロケーション（＋に取って代わる語、およびコロケーションタイプに応じた文字列の中の別の語によって形成される）を含まない候補が、ステップ２７８で削除される。残りの候補は、文字列候補と一致したクエリテンプレートの対応する重みに基づくスコアに応じて、ランク付けされる。例えば、クエリテンプレートの重みは、そのクエリテンプレートの中の語の数に基づくことが可能である。各候補に関するスコアは、それらの候補を含むすべての要約にわたって重みの合計をとることにより、計算される。候補を取得するクエリテンプレート（$ＱＴ_{S}$）に関するスコアは、以下によって表現することが可能である。すなわち、

【００３２】

【数１】

$$\text{スコア（候補）} = \sum_{QT_S} \text{重さ（ＱＴ）}$$

【００３３】

次に、ステップ２８０で、ランク付けされた候補リストがユーザに提示される。例えば、ポップアップ・メニューを使用して、ランク付けされたリストが提示されることが可能

である。ユーザは、そのリストから選択肢の１つを選び、コロケーション誤りを訂正することができる。

【００３４】

主題を、構造上の特徴および／または方法上の動作に特有の言い回しで説明してきたが、添付の特許請求の範囲において規定される主題は、必ずしも前述した特定の特徴または動作に限定されないことを理解されたい。むしろ、前述した特定の特徴および動作は、請求項を実施する例示的な形態として開示されている。

【図面の簡単な説明】

【００３５】

【図１】一般的なコンピューティング環境を示すブロック図である。

【図２】コロケーション誤りを検出して、訂正するためのシステムを示す流れ図である。

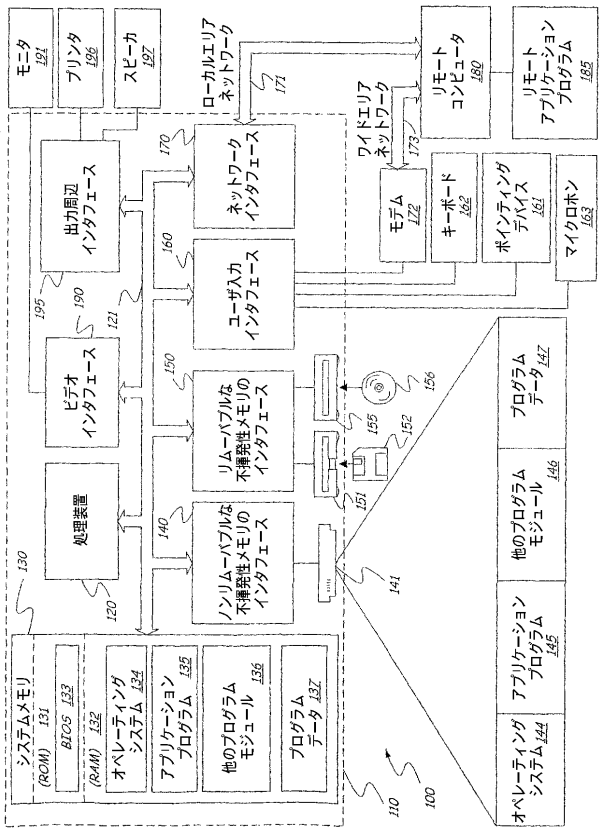
【図３】コロケーション誤りを検出して、訂正するための方法を示す流れ図である。

【図４】クエリ生成モジュールを示すブロック図である。

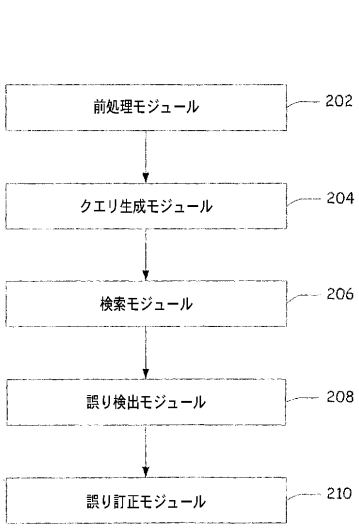
【図５】コロケーション誤りを検出するための方法を示す流れ図である。

【図６】候補コロケーション訂正を提示するための方法を示す流れ図である。

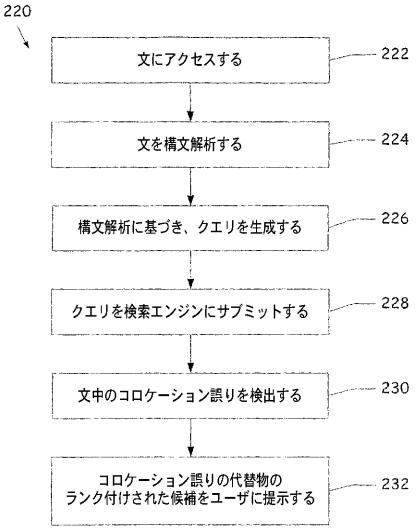
【図１】



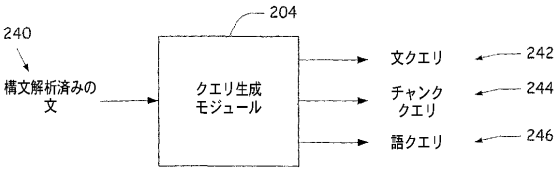
【図２】



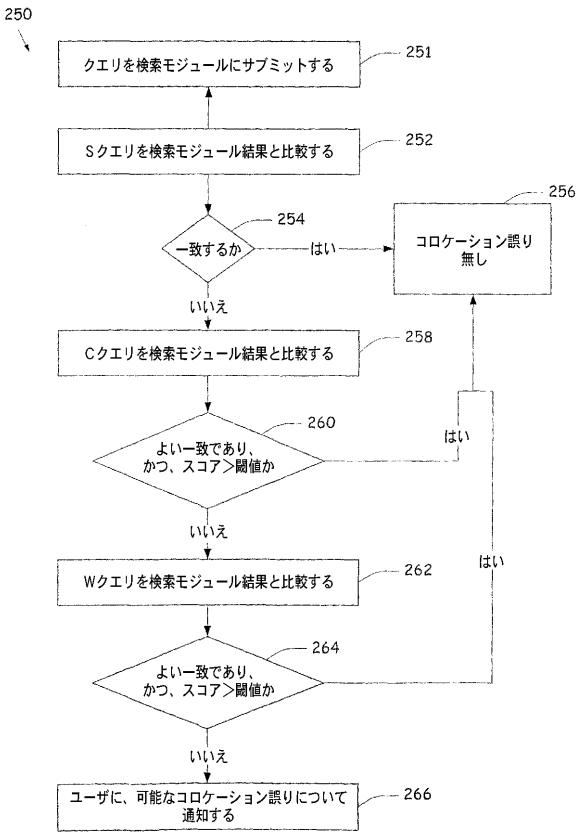
【図３】


220
文にアクセスする
222
文を構文解析する
224
構文解析に基づき、クエリを生成する
226
クエリを検索エンジンにサブミットする
228
文中のコロケーション誤りを検出する
230
コロケーション誤りの代替物の
ランク付けされた候補をユーザに提示する
232

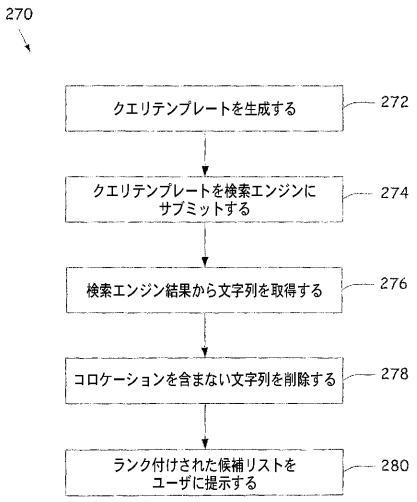
【図４】


240
構文解析済みの
文
204
クエリ生成
モジュール
文クエリ
242
チャンク
クエリ
244
語クエリ
246

【図５】


250
クエリを検索モジュールにサブミットする
251
Sクエリを検索モジュール結果と比較する
252
一致するか
254
はい
コロケーション誤り
無し
256
いいえ
Cクエリを検索モジュール結果と比較する
258
よい一致であり、
かつ、スコア＞閾値か
260
はい
いいえ
Wクエリを検索モジュール結果と比較する
262
はい
よい一致であり、
かつ、スコア＞閾値か
264
いいえ
ユーザに、可能なコロケーション誤りについて
通知する
266

【図６】


270
クエリテンプレートを生成する
272
クエリテンプレートを検索エンジンに
サブミットする
274
検索エンジン結果から文字列を取得する
276
コロケーションを含まない文字列を削除する
278
ランク付けされた候補リストを
ユーザに提示する
280

フロントページの続き

(72)発明者 シャオーウーエン ホン
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション インターナショナル パテンツ内
(72)発明者 チェンフェン ガオ
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション インターナショナル パテンツ内
(72)発明者 ミン チョウ
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ワン マイクロソフト ウェイ マイクロソフト コーポレーション インターナショナル パテンツ内

審査官 長 由紀子

(56)参考文献 特開２００１－１０１１８６（ＪＰ，Ａ）
大鹿 広憲 外3名，Ｇｏｏｇｌｅを活用した英作文支援システムの構築，ＤＥＷＳ２００５論文集 ［ｏｎｌｉｎｅ］，日本，（社）電子情報通信学会データ工学研究専門委員会，２００５年 5月 2日，DWS2005 4B-i8
大鹿 広憲 外3名，検索エンジンを使った翻訳サポートシステムの構築，情報処理学会研究報告，日本，社団法人情報処理学会，２００４年 7月１５日，第2004号第72号，p.585-591

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／２０－２８
Ｇ０６Ｆ １７／３０