(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4459443号
(P4459443)
(45)発行日　平成22年4月28日(2010.4.28)　(24)登録日　平成22年2月19日(2010.2.19)

(51)Int.Cl.　　F I
G06F 17/27 (2006.01)　　G06F 17/27　E
G06F 17/21 (2006.01)　　G06F 17/21　592J

請求項の数 5　(全 25 頁)

(21)出願番号	特願2000-551333 (P2000-551333)
(86)(22)出願日	平成11年5月28日 (1999.5.28)
(65)公表番号	特表2002-517039 (P2002-517039A)
(43)公表日	平成14年6月11日 (2002.6.11)
(86)国際出願番号	PCT/US1999/011856
(87)国際公開番号	WO1999/062001
(87)国際公開日	平成11年12月2日 (1999.12.2)
審査請求日	平成18年5月29日 (2006.5.29)
審判番号	不服2007-19673 (P2007-19673/J1)
審判請求日	平成19年7月12日 (2007.7.12)
(31)優先権主張番号	09/087,468
(32)優先日	平成10年5月29日 (1998.5.29)
(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)復代理人　100084191
弁理士　合田　潔
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(72)発明者　ウ　アンディー
アメリカ合衆国　９８００７　ワシントン州　ベルビュー　１５２　アベニュー　サウスイースト　２０３６

最終頁に続く

(54)【発明の名称】中国語テキストにおける単語分割

(57)【特許請求の範囲】
【請求項１】
コンピュータシステムにおいて中国語テキストセンテンス中に現れる個々の単語を識別するための単語識別方法であって、前記コンピュータシステムは、プログラムを記憶したメモリを有し、前記コンピュータシステムのＣＰＵが前記メモリに記憶されたプログラムに基づいて実行する単語識別方法であって、
複数の文字の各々について、
前記ＣＰＵが、前記文字で開始される単語の第２番目の位置に現れる文字の組の第１の指示を格納するステップと、
前記ＣＰＵが、単語の中で前記文字が現れる位置の組の第２の指示を格納するステップと、
前記センテンス中に現れる複数の連続した文字グループの各々について、
前記ＣＰＵが、前記第１の指示から前記グループの第２番目の位置に現れた文字が、前記グループの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されているかどうかを判定するステップと、
前記ＣＰＵが、もし、前記グループの第２番目の位置に現れた文字が、前記グループの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されていると判定された場合、前記第２の指示から前記グループの各々の文字が、前記グループ中に現れる位置において前記単語中に現れることを指示されているかどうかを判定するステップと、

前記ＣＰＵが、もし、前記グループの各々の文字が、前記グループ中に現れる位置において前記単語中に現れることを指示されていると判定された場合、文字の前記グループが中国語に使用されている単語であるかどうかを判定するために、文字の前記グループを中国語に使用されている単語のリストと比較するステップと、

前記ＣＰＵが、中国語に使用されている単語であると判定された文字の前記グループを構文解析パーサに提供するステップと、

前記ＣＰＵが、前記構文解析パーサから前記センテンスの構文構造を識別するパース構造を受け取るステップと、前記パース構造は提供された文字の前記グループのサブセットを前記センテンスの構文構造の一部として指定するものであり、

前記ＣＰＵが、前記パース構造により指定された文字の前記グループを、前記センテンス中に現れる単語として識別するステップと、

を含むことを特徴とする単語識別方法。

【請求項２】

請求項１に記載の単語識別方法において、前記ＣＰＵが、中国語に使用される単語のリストを読むステップと、前記ＣＰＵが、読まれた単語の前記リストから文字の出現に基づいて、前記格納された前記第１及び第２の指示を生成するステップとをさらに含むことを特徴とする単語識別方法。

【請求項３】

コンピュータシステムにおいて分断されない言語のテキストセンテンス中に現れる個々の単語を識別するための単語識別方法であって、前記コンピュータシステムは、プログラムを記憶したメモリを有し、前記コンピュータシステムのＣＰＵが前記メモリに記憶されたプログラムに基づいて実行する単語識別方法であって、

複数の文字の各々について、

前記ＣＰＵが、前記文字で開始される単語の第２番目の位置に現れる文字の組の第１の指示を格納するステップと、

前記ＣＰＵが、単語の中で前記文字が現れる位置の組の第２の指示を格納するステップと、

前記センテンス中に現れる複数の連続した文字グループの各々について、

前記ＣＰＵが、前記第１の指示から前記グループの第２番目の位置に現れた文字が、前記グループの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されているかどうかを判定するステップと、

前記ＣＰＵが、もし、前記グループの第２番目の位置に現れた文字が、前記グループの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されていると判定された場合、前記第２の指示から前記グループの各々の文字が、前記グループ中に現れる位置において前記単語中に現れることを指示されているかどうかを判定するステップと、

前記ＣＰＵが、もし、前記グループの各々の文字が、前記グループ中に現れる位置において前記単語中に現れることを指示されていると判定された場合、文字の前記グループが単語であるかどうかを判定するために文字の前記グループを単語のリストと比較するステップと、

前記ＣＰＵが、単語であると判定された文字の前記グループを構文解析パーサに提供するステップと、

前記ＣＰＵが、前記構文解析パーサから前記センテンスの構文構造を識別するパース構造を受け取るステップと、前記パース構造は提供された文字の前記グループのサブセットを前記センテンスの構文構造の一部として指定するものであり、

前記ＣＰＵが、前記パース構造により指定された文字の前記グループを、前記センテンス中に現れる単語として識別するステップと、

を含むことを特徴とする単語識別方法。

【請求項４】

コンピュータシステムにおいてテキストセンテンス中に現れる個々の単語を識別するた

めの単語識別方法であって、前記コンピュータシステムは、プログラムを記憶したメモリを有し、前記コンピュータシステムのＣＰＵが前記メモリに記憶されたプログラムに基づいて実行する単語識別方法であって、
複数の文字の各々について、
前記ＣＰＵが、前記文字で開始される単語の第２番目の位置に現れる文字の組の第１の指示を格納するステップと、
前記ＣＰＵが、単語の中で前記文字が現れる位置の組の第２の指示を格納するステップと、
前記センテンス中に現れる複数の連続した文字グループの各々について、
前記ＣＰＵが、前記第１の指示から前記グループの第２番目の位置に現れた文字が、前記グループの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されているかどうかを判定するステップと、
前記ＣＰＵが、もし、前記グループの第２番目の位置に現れた文字が、前記グループの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されていると判定された場合、前記第２の指示から前記グループの各々の文字が、前記グループ中に現れる位置において前記単語中に現れることを指示されているかどうかを判定するステップと、
前記ＣＰＵが、もし、前記グループの各々の文字が、前記グループ中に現れる位置において前記単語中に現れることを指示されていると判定された場合、文字の前記グループが単語であるかどうかを判定するために文字の前記グループを単語のリストと比較するステップと、
前記ＣＰＵが、単語であると判定された文字の前記グループを構文解析パーサに提供するステップと、
前記ＣＰＵが、前記構文解析パーサから前記センテンスの構文構造を識別するパース構造を受け取るステップと、前記パース構造は提供された文字の前記グループのサブセットを前記センテンスの構文構造の一部として指定するものであり、
前記ＣＰＵが、前記パース構造により指定された文字の前記グループを、前記センテンス中に現れる単語として識別するステップと
を含むことを特徴とする単語識別方法。

【請求項５】

コンピュータシステムにおいて自然言語文字のセンテンスから単語である可能性の文字の組合わせを選択する方法であって、複数の文字の各々について、前記文字で開始される単語の第２番目の位置に現れる文字の指示及び単語の中で前記文字が現れる位置の指示を使用して選択する方法であって、前記コンピュータシステムは、プログラムを記憶したメモリを有し、前記コンピュータシステムのＣＰＵが前記メモリに記憶されたプログラムに基づいて実行する単語選択方法であって、
前記センテンス中に現れる複数の連続した文字の組合わせの各々について、
前記ＣＰＵが、前記組合わせの第１番目の位置に現れる文字で開始される単語の第２番目の位置に現れる文字の組を示す記憶された第１の指示をアクセスすることにより、前記組合わせの第２番目の位置に現れた文字が、前記組合わせの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されているかどうかを判定するステップと、
前記ＣＰＵが、もし、前記組合わせの第２番目の位置に現れた文字が、前記組合わせの第１番目の位置に現れた前記文字で開始される単語中に現れることを指示されていると判定された場合、単語の中で各々の文字が現れる位置の組を示す記憶された第２の指示をアクセスすることにより、前記組合わせの各々の文字が、前記組合わせ中に現れる位置において前記単語中に現れることを指示されているかどうかを判定するステップと、
前記ＣＰＵが、もし、前記組合わせの各々の文字が、前記組合わせ中に現れる位置において前記単語中に現れることを指示されていると判定された場合、文字の前記組合わせは単語である可能性があると判定するステップと、

を含むことを特徴とする単語選別方法。

【発明の詳細な説明】

【０００１】

（技術分野）

本発明は、一般に自然言語処理の分野に関し、より詳細には、単語分割（ｗｏｒｄ　ｓｅｇｍｅｎｔａｔｉｏｎ）の分野に関する。

【０００２】

（発明の背景）

単語分割とは、テキストなどの言語表現を構成する個々の単語を識別するプロセスのことである。単語分割は、すべて個々の単語の識別の恩恵を受ける、綴りおよび文法の検査、テキストからの話法の合成、および自然言語の解剖および理解の実行に有用である。

【０００３】

英語テキストの単語分割の実行は、空白および句読点が一般にテキスト中の個々の単語を区切るのでかなり簡単である。下記の表１の英語センテンスについて考えてみる。

【０００４】

【表１】

The motion was then tabled--that is, removed indefinitely from consideration.

表１

【０００５】

表１の英語センテンスは、空白および／または句読点の連続した各シーケンスをそのシーケンスの前にある単語の末尾として識別することによって、下記の表２に示すように簡単に分割することができる。

【０００６】

【表２】

The motion was then tabled--that is, removed indefinitely from consideration.

表２

【０００７】

中国語テキストでは、単語の境界は明示的ではなく暗示的である。「委員会は昨日の午後ブエノスアイレスでこの問題について議論した。」を意味する、下記の表３のセンテンスについて考えてみる。

【０００８】

【表３】

昨天下午委员会在布宜诺斯艾利斯讨论了这个问题。

表３

【０００９】

センテンスに句読点および空白がないにもかかわらず、中国語の読者は、表３のセンテンスを、下記の表４で別々に下線を付けられた単語から構成されるものとして認識するであろう。

【００１０】

【表４】
昨天下午委员会在布宜诺斯艾利斯讨论了这个问题。

表４

【００１１】
上記の例から、中国語の単語分割は英語の単語分割と同じ形では実行できないことが分かる。とは言え、中国語分割を自動的に実行する正確で効率的な手法には、かなりの効用があろう。

【００１２】
（発明の概要）
本発明は、単語であることがある文字の自然言語文字組合せのシーケンスから選択するための機能（ｆａｃｉｌｉｔｙ）を提供する。この機能は複数の単語の各々についての隣接する文字に応じた確率の指示を使用する。

【００１３】
本発明の一態様は、テキストセンテンス中に現れる個々の単語を識別するためのコンピュータシステムにおける方法である。この方法は、複数の単語の各々について、その単語が自然言語テキスト中に現れるかどうかの、隣接する文字に応じた確率の指示を格納するステップと、センテンス中に現れる連続した複数の文字グループの各々について、重なる可能性のある単語を判定するステップと、格納された指示および隣接する文字に基づいて確率を確認するステップと、可能性のある単語であると判定された文字グループを確率の指示とともにパーサ（ｐａｒｓｅｒ）に提出するステップとを含む。この方法を実現する命令を格納するためのコンピュータ可読媒体も提供される。

【００１４】
本発明の第２の態様は、自然言語テキスト中に現れる個々の単語を識別するのに使用する単語分割データ構造を含有するコンピュータメモリを含む。このデータ構造は、複数の単語の各々について、その単語が自然言語テキスト中に現れるかどうかの、隣接する文字に応じた確率の指示を含む。

【００１５】
（発明の詳細な説明）
本発明は、中国語テキストにおける単語分割を提供する。好ましい実施形態では、単語分割ソフトウェア機能（「機能」）は、（１）入力センテンス中の文字の可能性のある組合せを評価し、入力センテンス中の単語を表す可能性のない組合せを廃棄すること、（２）辞書中の文字の残りの組合せを調べて、それらが単語を構成することができるかどうかを判定すること、（３）単語であると判定された文字の組合せを、入力センテンスを表す代替単語彙記録（lexical record）として自然言語パーサに提出することによって、中国語などの非分割言語におけるテキストのための単語分割を提供する。パーサは入力センテンスの構文構造（syntactic structure）を表す構文パースツリー（syntactic parse tree）を生成し、パースツリーは、入力センテンス中の単語であることが証明された文字の組合せを表す語彙記録のみを含有する。語彙記録をパーサに提出するとき、文字のより短い組合せよりもセンテンスの正確な分割をより一般的に表す文字のより長い組合せが、文字のより短い組合せの前にパーサによって考慮されるように、機能は語彙記録を重み付けする。

【００１６】
入力センテンス中の単語を表す可能性のない文字の組合せを廃棄することを容易にするために、機能は、辞書中に現れる各文字について、（１）単語が現れる単語長および文字位置のすべての異なる組合せの指示、および（２）この文字が単語を開始するときにこの文字の後にくることができるすべての文字の指示を、辞書に追加する。機能はさらに、（３）複数文字語内の下位単語が独立可能であり、考慮すべきかどうかの指示を複数文字語に追加する。センテンスを処理する場合、機能は、（１）辞書中に現れない単語長／位置組合せ中に使用されている文字の組合せ、および（２）第２の文字が第１の文字の第２の可

能性のある文字として掲載されていない文字の組合せを廃棄する。機能はさらに、（３）下位単語を考慮すべきでない単語中に現れる文字の組合せを廃棄する。

【００１７】

このようにして、機能は、辞書中で調べられた文字組合せの数を最小にするだけでなく、センテンスの構文文脈を利用して、それぞれ有効な単語から構成される代替分割結果どうしを区別する。

【００１８】

図１は、機能が実行することが好ましい汎用コンピュータシステムの高レベルブロック図である。コンピュータシステム１００は中央処理ユニット（ＣＰＵ）１１０、入出力デバイス１２０、およびコンピュータメモリ（メモリ）１３０を含有する。入出力デバイスとしては、ハードディスクドライブなどの記憶デバイス１２１、コンピュータ可読媒体上に設けられる機能を含む、ソフトウェア製品をインストールするために使用することができるコンピュータ可読媒体ドライブ１２２、コンピュータシステム１００が他の接続されたコンピュータシステム（図示せず）と通信することができるネットワーク接続１２３がある。メモリ１３０は、中国語テキスト中に現れる個々の単語を識別するための単語分割機能１３１、自然言語テキスト中に現れる単語を表す語彙記録から自然言語テキストセンテンスの構文構造を表すパースツリーを生成するための構文パーサ１３３、およびパースツリー用の語彙記録を構成する際にパーサによって使用され、また自然言語テキスト中に現れる単語を識別するために機能によって使用される語彙知識ベース１３２を含有することが好ましい。機能は上述のように構成されるコンピュータシステム上で実現することが好ましいが、当業者であれば異なる構成を有するコンピュータシステム上で実現することもできることを理解できよう。

【００１９】

図２は、機能が動作することが好ましい２つの段階を示す概略流れ図である。ステップ２０１で、初期化段階の一部として、機能は、単語分割を実行するために機能によって使用される情報を含むように語彙知識ベースを増大させる。ステップ２０１については図３に関して以下でより詳細に議論する。手短に言えば、ステップ２０１で、機能は、語彙知識ベース中の任意の単語中に現れる文字の項目を語彙知識ベースに追加する。各文字の追加された項目は、文字が単語中に現れる異なる位置を指示するＣｈａｒＰｏｓ属性を含む。各文字の項目はさらに、現在の文字で始まる単語の第２の位置に現れる文字の集合を指示するＮｅｘｔＣｈａｒｓ属性を含有する。最後に、機能はまた、現在の単語を一緒に構成するより小さい単語を含むように、単語を含んでいる文字のシーケンスをさらに考慮すべきかどうかを指示する語彙知識ベース中に現れるＩｇｎｏｒｅＰａｒｔｓ属性を各単語に追加する。

【００２０】

ステップ２０１の後、機能はステップ２０２に進み、初期化段階を終了し、単語分割段階を開始する。単語分割段階では、機能は、語彙知識ベースに追加された情報を使用して、中国語テキストセンテンスの単語分割を実行する。ステップ２０２で、機能は単語分割のための中国語テキストセンテンスを受け付ける。ステップ２０３で、機能は受け付けたセンテンスをその成分語に分割する。ステップ２０３については図５に関して以下でより詳細に議論する。手短に言えば、機能は、センテンス中の文字の可能性のあるすべての連続した組合せの小部分を語彙知識ベース中で調べる。次いで機能は、語彙知識ベースによって単語であることが指示された調べた文字の組合せを構文パーサに提出する。パーサは、センテンスの構文構造を判定する場合、その著者によってセンテンス中の単語を含むように意図された文字の組合せを識別する。ステップ２０３の後、機能はステップ２０２に進み、単語分割のための次のセンテンスを受け付ける。

【００２１】

図３は、単語分割を実行するために使用される情報を含むように初期化段階において語彙知識ベースを増大させるために機能によって実行されることが好ましいステップを示す流れ図である。これらのステップは、（ａ）語彙知識ベース中の単語中に現れる文字の項目

を語彙知識ベースに追加し、（ｂ）語彙知識ベース中の文字項目にＣｈａｒＰｏｓおよびＮｅｘｔＣｈａｒｓ属性を追加し、（ｃ）語彙知識ベース中の単語の項目にＩｇｎｏｒｅＰａｒｔｓ属性を追加する。

【００２２】

ステップ３０１～３１２で、機能は語彙知識ベース中の各単語項目をループする。ステップ３０２で、機能は単語中の各文字位置をループする。すなわち、３つの文字を含有する単語の場合、機能は単語の第１、第２、および第３の文字をループする。ステップ３０３で、現在の文字位置にある文字が語彙知識ベース中に項目を有する場合、機能はステップ３０５に進み、そうでない場合、機能はステップ３０４に進む。ステップ３０４で、機能は現在の文字の項目を語彙知識ベースに追加する。ステップ３０４の後、機能は３０５に進む。ステップ３０５で、機能は、その文字が現在の単語中に現れている位置にその文字が現れることができることを指示する順序付けされた対を、語彙知識ベース中の文字の項目中に格納されたＣｈａｒＰｏｓ属性に追加する。追加された順序付けされた対は、形式（位置、長さ）を有する。ここで、位置は文字が単語中で占有している位置であり、長さは単語中の文字の数である。例えば、単語

【００２３】

【外１】

「委员会」

【００２４】

中の文字「委」について、機能は、文字「委」の語彙知識ベース項目中のＣｈａｒＰｏｓ属性中に格納された順序付けされた対のリストに、順序付けされた対（１、３）を追加することになる。順序付けされた対がすでに現在の単語のＣｈａｒＰｏｓ属性中に含有されている場合、機能は、ステップ３０５で説明したように順序付けされた対を追加しないことが好ましい。ステップ３０６で、処理すべき追加の文字が現在の単語中に残っている場合、機能はステップ３０２に進み、次の文字を処理し、そうでない場合、機能はステップ３０７に進む。

【００２５】

ステップ３０７で、単語が単数文字語である場合、機能はステップ３０９に進み、そうでない場合、機能はステップ３０８に進む。ステップ３０８で、機能は、現在の単語の位置にある文字を、現在の単語の第１の位置にある文字の語彙知識ベース記録中のＮｅｘｔＣｈａｒｓ属性中の文字のリストに追加する。例えば、単語

【００２６】

【外２】

「委员会」

【００２７】

の場合、機能は文字

【００２８】

【外３】

「员」

【００２９】

のＮｅｘｔＣｈａｒｓ属性の格納された文字のリストに文字「委」を追加する。ステップ３０８の後、機能は３０９に進む。

【００３０】

ステップ３０９で、現在の単語が他のより小さい単語を含有することができる場合、機能はステップ３１１に進み、そうでない場合、機能はステップ３１０に進む。ステップ３０９については図４に関して以下でより詳細に議論する。手短に言えば、機能は、いくつかのヒューリスティックス（heuristics）を使用して、現在の単語を構成する文字のシーケ

ンスの発生がある文脈では２つ文字それ以上のより小さい単語を構成することができるかどうかを判定する。

【００３１】

ステップ３１０で、機能は単語の語彙知識ベース項目中に単語のＩｇｎｏｒｅＰａｒｔｓ属性をセットする。ＩｇｎｏｒｅＰａｒｔｓ属性をセットすることは、機能が入力テキストセンテンス中でこの単語に遭遇したときに、この単語がより小さい単語を含有するかどうかを判定するためにさらなるステップを実行すべきであることを指示する。ステップ３１０の後、機能は３１２に進む。ステップ３１１で、現在の単語は他の単語を含有することができるので、機能は単語のＩｇｎｏｒｅＰａｒｔｓ属性をクリアし、それにより機能は、入力テキストセンテンス中でこの単語に遭遇したときに、単語がより小さい単語を含有するかどうかの調査に取りかかる。ステップ３１１の後、機能は３１２に進む。ステップ３１２で、処理すべき追加の文字が語彙知識ベース中に残っている場合、機能はステップ３０１に進み、次の単語を処理し、そうでない場合、これらのステップは終了する。

【００３２】

機能は、各文字にＣｈａｒＰｏｓおよびＮｅｘｔＣｈａｒ属性を割り振ることによって語彙知識ベースを増大させるために図３に示されるステップを実行するとき、以下の表５に示すように、表３に示されるサンプルセンテンス中に現れる文字にこれらの属性を割り振る。

【００３３】

【表５】

文字	CharPos	NextChar
昨	(1, 2) (1, 3) (3, 4)	儿天晚
天	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4)	安崩兵
下	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4)	巴百班…午…
午	(1, 2) (2, 2) (2, 3) (2, 4)	餐饭后…
委	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (3, 4) (4, 4) (3, 5)	靡派屈…员…
员	(1, 2) (2, 2) (2, 3) (3, 3) (2, 4) (3, 4) (4, 4)	额工司外
会	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4) (4, 5)	标餐操…
在	(1, 2) (2, 2) (1, 3) (2, 3) (1, 4) (2, 4) (3, 4) (4, 4)	案场朝…
布	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4) (1, 5) (2, 5) (3, 5) (4, 5) (1, 6) (2, 6) (1, 7)	达店丁…宜…
宜	(1, 2) (2, 2) (2, 3) (3, 3) (2, 4) (3, 4) (4, 4) (3, 6) (2, 7)	宾昌城…
诺	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (2, 4) (3, 4) (4, 4) (3, 7)	贝丁曼萨言
斯	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4) (1, 5) (2, 5) (3, 5) (4, 5) (5, 5) (1, 6) (3, 6) (4, 6) (5, 6) (6, 6) (4, 7) (5, 7) (6, 7) (7, 7)	文德拉…
艾	(1, 2) (2, 2) (1, 3) (3, 4) (4, 4) (1, 5) (5, 7)	比丁黎…
利	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (2, 5) (3, 5) (4, 5) (5, 6) (6, 7)	薄比弊…
斯	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4) (1, 5) (2, 5) (3, 5) (4, 5) (5, 5) (1, 6) (3, 6) (4, 6) (5, 6) (6, 6) (4, 7) (5, 7) (6, 7) (7, 7)	文德拉…
讨	(1, 2) (2, 2) (1, 3) (2, 3) (1, 4) (2, 4)	伐饭好价叫论人厌
论	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4)	处点调…
了	(1, 2) (2, 2) (1, 3) (3, 3) (1, 4) (2, 4) (3, 4) (4, 4)	不结解…
这	(1, 2) (1, 3) (1, 4)	边儿个…
个	(1, 2) (2, 2) (1, 3) (2, 3) (3, 3) (1, 4) (2, 4)	别儿旧…
问	(1, 2) (2, 2) (1, 3) (2, 3) (1, 4) (3, 4) (4, 4)	长答道…题…
题	(1, 2) (2, 2) (2, 3) (3, 3) (2, 4) (4, 4)	跋材词…

表５：文字語彙知識ベース項目

【００３４】

表５から、例えば、文字「昨」のＣｈａｒＰｏｓ属性から、この文字は、長さが２文字、３文字、または４文字の単語の第１の文字のように見えることが分かる。さらに、文字「昨」のＮｅｘｔＣｈａｒ属性から、この文字から始まる単語において、第２の文字は「儿」、「天」または「晚」のいずれかであることが分かる。

【００３５】

図４は、特定の単語が他のより小さい単語を含有することができるかどうかを判定するた

めに実行されることが好ましいステップを示す流れ図である。英語との類推として、空白および句読点文字を英語センテンスから取り除いた場合、文字のシーケンス「ｂｅａｔ」は単語「ｂｅａｔ」かまたは２つの単語「ｂｅ」および「ａｔ」と解釈することができよう。ステップ４０１で、単語が４つまたはそれ以上の文字を含有する場合、機能はステップ４０２に進み、単語が他の単語を含有することができないという結果を戻し、そうでない場合、機能はステップ４０３に進む。ステップ４０３で、単語中のすべての文字が単数文字語を構成することができる場合、機能はステップ４０５に進み、そうでない場合、機能はステップ４０４に進み、単語が他の単語を含有することができないという結果を戻す。ステップ４０５で、単語が派生接辞、すなわち接頭辞または接尾辞として頻繁に使用される単語を含有する場合、機能はステップ４０６に進み、単語が他の単語を含有することができないという結果を戻し、そうでない場合、機能はステップ４０７に進む。ステップ４０７で、単語中の隣接する文字対が、言語のテキスト中に隣接して現れたときにしばしば分割される場合、機能はステップ４０９に進み、単語が他の単語を含有することができるという結果を戻し、そうでない場合、機能はステップ４０８に進み、単語が他の単語を含有することができないという結果を戻す。

【００３６】

特定の単語が他のより小さい単語を含有することができるかどうかを判定した結果を以下の表６に示す。

【００３７】

【表６】

単語	IgnoreParts
昨天	セット
天下	クリア
下午	セット
委员会	クリア
委员	セット
布宜诺斯艾利斯	セット
讨论	セット
这个	クリア
问题	セット

表６：単語語彙知識ベース項目

【００３８】

例えば、表６から、機能は、単語「昨天」が他の単語を含有することができないが、単語「天下」が他の単語を含有することができると判定したことが分かる。

【００３９】

図５は、センテンスをその成分語に分割するために機能によって実行されることが好ましいステップの流れ図である。これらのステップは、センテンス中に現れる言語の異なる単語を識別する単語リストを生成する。次いで単語リストは、その著者によってセンテンス中の単語を含むように意図された単語リスト中の部分集合を識別するためにパーサに提出される。

【００４０】

ステップ５０１で、機能は、センテンス中に現れる複数文字語を単語リストに追加する。ステップ５０１については図６に関して以下でより詳細に議論する。ステップ５０２で、機能は、センテンス中に現れる単数文字語を単語リストに追加する。ステップ５０２については図９に関して以下でより詳細に議論する。ステップ５０３で、機能は、ステップ５０１および５０２で単語リストに追加された単語の語彙パーサによって使用される語彙記録を生成する。ステップ５０４で、機能は語彙記録に確率を割り振る。語彙記録の確率は

、語彙記録がセンテンスの正確なパースツリーの一部となる可能性を反映し、またパースプロセスにおける語彙記録の適用を順序付けするためにパーサによって使用される。パーサは、パースプロセス中に語彙記録をその確率の減少順に適用する。ステップ５０４については図１０に関して以下でより詳細に議論する。ステップ５０５で、機能は構文パーサを使用して、語彙記録を解剖して、センテンスの構文構造を反映するパースツリーをつくり出す。このパースツリーは、そのリーフとしてステップ５０３で生成された語彙記録の部分集合を有する。ステップ５０６で、機能は、パースツリーのリーフである語彙記録によって表される単語をセンテンスの単語として識別する。ステップ５０６の後、これらのステップは終了する。

【００４１】

図６は、単語リストに複数文字語を追加するために機能によって実行されることが好ましいステップの流れ図である。これらのステップは、複数文字語を識別するためにセンテンスを分析する際にセンテンス内の現在位置を使用する。これらのステップはさらに、図４に示される機能によって語彙知識ベースに追加されたＣｈａｒＰｏｓ、ＮｅｘｔＣｈａｒ、およびＩｇｎｏｒｅＰａｒｔｓ属性を利用する。第１の好ましい実施形態によれば、図６に示されるステップの実行中に必要に応じて語彙知識ベースからこれらの属性を取り出す。第２の好ましい実施形態では、センテンス中の文字のＮｅｘｔＣｈａｒ属性および／またはＣｈａｒＰｏｓ属性の値が第２のステップの実行の前に事前ロードされる。第２の好ましい実施形態に関して、センテンス中に現れる各文字のＣｈａｒＰｏｓ属性の値を含有する３次元アレイがメモリ中に格納されることが好ましい。このアレイは、センテンス中の所与の位置にある文字について、その文字が所与の長さの単語中の所与の位置にあることができるかどうかを指示する。これらの属性の値を捕らえることにより、図６に示されるステップを実行するときにそれらの値に公式にアクセスすることが可能となる。

【００４２】

ステップ６０１で、機能は、この位置を文章の最初の文字にセットする。ステップ６０２～６１４で、位置が文章の末尾に進むまでステップ６０３～６１３を継続して繰り返す。

【００４３】

ステップ６０３～６０９で、機能は、現行位置から始まる各単語候補を通ってループする。好ましくは、現行位置からスタートする７文字長の単語候補から始まり、各反復において、単語候補が２文字長になるまで単語候補の末尾から１文字を除去する。現行位置から始まる文章に残っている文字が７文字よりも少ない場合は、文章に残っている文字が不十分である単語候補に関する反復を機能が省略することが好ましい。ステップ６０４で、機能は、単語候補を構成している文字のＮｅｘｔＣｈａｒおよびＣｈａｒＰｏｓに関係する現行単語候補条件に関してテストする。ステップ６０４は、図７に関連して以下により詳細に論ずる。単語候補に関してＮｅｘｔＣｈａｒ条件とＣｈａｒＰｏｓ条件がどちらも満足されている場合は、機能はステップ６０５に続き、そうでない場合はステップ６０９に続く。ステップ６０５で、機能は、字句知識ベース中で単語候補を検索して、単語候補が単語であるかどうか判定する。ステップ６０６で、単語候補が単語である場合は、機能はステップ６０７に続き、そうでない場合はステップ６０９に続く。ステップ６０７で、機能は、単語候補を文章中に現れる単語のリストに追加する。ステップ６０８で、単語候補が他の単語を含む可能性がある場合、すなわち単語に関するＩｇｎｏｒｅＰａｒｔｓ属性がクリア状態である場合は、機能はステップ６０９に続き、そうでない場合はステップ６１１に続く。ステップ６０９で、処理すべきさらなる単語候補が残っている場合は、機能はステップ６０３に続き、次の単語候補を処理し、そうでない場合はステップ６１０に続く。ステップ６１０で、現行位置を文章の末尾に向けて一文字進める。ステップ６１０の後、機能はステップ６１４に続く。

【００４４】

ステップ６１１で、単語候補の最後の文字が、これも単語である可能性がある別の単語候補と重複している場合は、機能はステップ６１３に続き、そうでない場合はステップ６１２に続く。ステップ６１１は、図８に関連して以下により詳細に論ずる。ステップ６１２

で、機能は、単語候補の最後の文字の後にある文章中の文字に位置を進める。ステップ６１２の後、機能はステップ６１４に続く。ステップ６１３では、現行単語候補の最後の文字に位置を進める。ステップ６１３の後、機能はステップ６１４に続く。ステップ６１４で、位置が文章の末尾にない場合は、機能はステップ６０２に続いて、単語候補の新たなグループを考察し、そうでない場合はこれらのステップが終了する。

【００４５】

図７は、ある単語候補に関するＮｅｘｔＣｈａｒ条件およびＣｈａｒＰｏｓ条件をテストするために機能によって実施されることが好ましいステップを示す流れ図である。ステップ７０１で、単語候補の２番目の文字が、単語候補の最初の文字のＮｅｘｔＣｈａｒリスト中にある場合は、機能はステップ７０３に続き、そうでない場合はステップ７０２に続き、条件がどちらも満足されているという結果を返す。ステップ７０３～７０６で、機能は、単語候補中の各文字位置を通ってループする。ステップ７０４で、現行位置と単語候補の長さとから構成された順序対が、現行文字位置にある文字に関するＣｈａｒＰｏｓリスト中の順序対の中にある場合は、機能はステップ７０６に続き、そうでない場合はステップ７０５に続いて、条件が両方とも満足されているわけではないという結果を返す。ステップ７０６で、処理すべきさらなる文字位置が単語候補中に残っている場合は、機能はステップ７０３に続き、単語候補中の次の文字位置を処理し、そうでない場合はステップ７０７に続き、単語候補によって条件がどちらも満足されているという結果を返す。

【００４６】

図８は、現行単語候補の最後の文字が、単語である可能性がある別の単語候補と重複しているかどうかを判定するために機能によって実施されることが好ましいステップを示す流れ図である。ステップ８０１で、単語候補の後の文字が、単語候補の最後の文字に関するＮｅｘｔＣｈａｒ属性をもつ文字のリスト中にある場合は、機能はステップ８０３に続き、そうでない場合はステップ８０２に続き、重複がないという結果を返す。ステップ８０３で、機能は、最後の文字がない状態での単語候補を字句知識ベース中で検索して、最後の文字がない状態での単語候補が単語であるかどうか判定する。ステップ８０４で、最後の文字がない状態での単語候補が単語である場合は、機能はステップ８０６に続き、重複があるという結果を返し、そうでない場合はステップ８０５に続き、重複がないという結果を返す。

【００４７】

図６に示されたステップの動作の例示を下記の表７に示す。

【００４８】

【表７－１】

番号	組合せ	CharPos	NextChars	検索するか	単語か
1	昨天下午委员会	昨で否	合	いいえ	いいえ
2	昨天下午委员	昨で否	合	はい	はい
3	昨天下午委	昨で否	合	はい	はい
4	昨天下午	昨で否	否	いいえ	いいえ
5	昨天下	合	否	いいえ	いいえ
6	昨天	合	否	いいえ	いいえ
7	天下午委员会在	天で否	否	いいえ	いいえ
8	天下午委员会	天で否	否	いいえ	いいえ
9	天下午委员	天で否	否	いいえ	いいえ
１０	天下午委	午で否	否	いいえ	いいえ
１１	天下午	午で否	否	いいえ	いいえ
１２	天下	合	否	いいえ	いいえ
１３	下午委员会在布	下で否	否	いいえ	いいえ
１４	下午委员会在	下で否	否	いいえ	いいえ
１５	下午委员会	下で否	否	いいえ	いいえ
１６	下午委员	合	否	いいえ	いいえ
１７	下午委	合	否	いいえ	いいえ
１８	下午	合	否	いいえ	いいえ
１９	委员会在布宜诺	委で否	否	いいえ	いいえ
２０	委员会在布宜	委で否	否	いいえ	いいえ
２１	委员会在布	委で否	否	いいえ	いいえ
２２	委员会讨	讨で否	否	いいえ	いいえ
２３	委员会	合	否	いいえ	いいえ
２４	委员	合	否	いいえ	いいえ
２５	会在布宜诺斯艾	会で否	否	いいえ	いいえ
２６	会在布宜诺斯	会で否	否	いいえ	いいえ
２７	会在布宜诺	会で否	否	いいえ	いいえ
２８	会在布宜	合	否	いいえ	いいえ
２９	会在布	合	否	いいえ	いいえ
３０	会在	合	否	いいえ	いいえ

【００４９】
【表７－２】

３１	在布宜诺斯艾利	在で否	否	いいえ	いいえ
３２	在布宜诺斯艾	在で否	否	いいえ	いいえ
３３	在布宜诺斯	在で否	否	いいえ	いいえ
３４	在布宜诺	合	否	いいえ	いいえ
３５	在布宜	合	否	いいえ	いいえ
３６	在布	合	否	いいえ	いいえ
３７	布宜诺斯艾利斯	合	合	はい	はい
３８	讨论了这个问题	讨で否	合	いいえ	いいえ
３９	讨论了这个问	讨で否	合	いいえ	いいえ
４０	讨论了这个	讨で否	合	いいえ	いいえ
４１	讨论了这	这で否	合	いいえ	いいえ
４２	讨论了	合	合	はい	いいえ
４３	讨论	合	合	はい	はい
４４	了这个问题	了で否	否	いいえ	いいえ
４５	了这个问	这で否	否	いいえ	いいえ
４６	了这个	这で否	否	いいえ	いいえ
４７	了这	这で否	否	いいえ	いいえ
４８	这个问题	合	合	はい	いいえ
４９	这个问	问で否	合	いいえ	いいえ
５０	这个	合	合	はい	はい
５１	个问题	合	否	いいえ	いいえ
５２	个问	合	否	いいえ	いいえ
５３	问题	合	合	はい	はい

表７：考察される文字組合せ

【００５０】

表７は、機能によって考察されるサンプル文章からの文字の５３通りの組合せそれぞれに関して、ＣｈａｒＰｏｓテストの結果、ＮｅｘｔＣｈａｒｓテストの結果、機能が字句知識ベース中で単語を検索したかどうか、および文字の組合せが単語であることを字句知識ベースが示したかどうかを示す。

【００５１】

文字「昨」のＣｈａｒＰｏｓ属性が順序対（１，７）、（１，６）、（１，５）、または（１，４）を含まないため、組合せ１～４がＣｈａｒＰｏｓテストに失敗したことがわかる。一方、組合せ５および６では、ＣｈａｒＰｏｓテストとＮｅｘｔＣｈａｒｓテストがどちらも合格されている。したがって、機能が字句知識ベース中で組合せ５および６を検索し、組合せ５は単語でなく、組合せ６は単語であると判定する。組合せ６を処理し、現行位置をどのくらい遠くに進めるかを決定した後、機能は、ＩｇｎｏｒｅＰａｒｔｓ属性がセット状態である、しかし単語「昨天」が、文字「天」から始まる単語候補と重複していると判定する。したがって、機能は、ステップ６１３に従って組合せ６の末尾にある文字「天」に進む。組合せ７～１２では、組合せ１２のみがＣｈａｒＰｏｓテストとＮｅｘｔＣｈａｒｓテストに合格する。したがって、組合せ１２が検索され、単語であると判定される。組合せ１２を処理し、現行位置をどのくらい遠くに進めるかを決定した後、機能は、組合せ１２によって構成される単語のＩｇｎｏｒｅＰａｒｔｓ属性がクリア状態であると判定し、したがって現行位置を、組合せ１２に後続する文字にではなく文字「下」に１文字進める。

【００５２】

さらに、組合せ１８、２４、３７、４３は、それらのＩｇｎｏｒｅＰａｒｔｓ属性がセット状態であり、それらの最終文字において、単語である可能性がある任意の単語候補と重

複していない単語であることがわかる。したがって、それぞれを処理した後、機能は、ステップ６１２に従ってその文字組合せに後続する文字に現行位置を進め、それによりこれら４つの組合せそれぞれに関する最大４１通りのさらなる組み合わせを不必要に処理することを省略する。

【００５３】

さらに、組合せ２３および５０によって構成された単語のＩｇｎｏｒｅＰａｒｔｓ属性は、クリア状態であることがわかる。そのため機能は、これらの組合せを処理した後、ステップ６１０に従って１文字だけ現行位置を進める。

【００５４】

さらに、２文字組合せ３０、３６、４７、および５２は、機能によって単語を構成すると判定されないことがわかる。したがって、機能は、ステップ６１０に従ってこれらの組合せを処理した後、１文字だけ現行位置を進める。全体として、機能は、サンプル文章における１１２通りの可能な組合せのうち１４通りのみを検索する。機能によって検索された１４通りの組合せうち、９通りが実際に真の単語である。

【００５５】

以下表８に示されるように、表７に関連して記述された処理の後、単語リストは、組合せ６、１２、１８、２３、２４、３７、４３、５０、および５３によって構成される単語を含む。

【００５６】

【表８】

番号	単語	品詞
６	昨天	名詞
１２	天下	名詞
１８	下午	名詞
２４	委员	名詞
２３	委员会	名詞
３７	布宜诺斯艾利斯	名詞
４３	讨论	動詞
５０	这个	代名詞
５３	问题	名詞

表８：複数文字単語を有する単語リスト

【００５７】

図９は、単一文字単語を単語リストに追加するために機能によって実施されることが好ましいステップを示す流れ図である。ステップ９０１～９０６で、機能は、最初の文字から最後の文字まで文章中の各文字を通ってループする。ステップ９０２で、機能は、字句知識ベース中のそのエントリに基づいて、文字が単一文字単語を構成するかどうか判定し、そうでない場合は、文字を単語リストに追加することなくステップ９０６に続く。文字が単一文字単語を構成する場合は、機能がステップ９０３に続き、そうでない場合は、文字を単語リストに追加することなくステップ９０６に続く。ステップ９０３で、他の単語を含まない可能性がある単語中に文字が含まれている場合、すなわち単語リストにすでにある単語がそのＩｇｎｏｒｅＰａｒｔｓ属性をセット状態で有している場合、機能はステップ９０４に続き、そうでない場合はステップ９０５に続いて、文字を単語リストに追加する。ステップ９０４で、単語リストにある他の単語と重なる単語リストにある単語中に文字が含まれている場合は、機能は、文字を単語リストに追加することなくステップ９０６に続き、そうでない場合はステップ９０５に続く。ステップ９０５で、機能は、現行文字を備える単一文字単語を単語リストに追加する。ステップ９０６で、処理すべきさらなる文字が文章中に残っている場合は、機能はステップ９０１に続き、文章中の次の文字を処

理し、そうでない場合はこれらのステップが終了する。

【００５８】

以下の表９は、図９に示されるステップを実施したときに、機能が単一文字単語５４～６１を単語リストに追加することを示す。

【００５９】

【表９】

番号	単語	品詞
６	昨天	名詞
５４	昨	形態素
５５	天	名詞
１２	天下	名詞
５６	下	名詞（ローカライザ）
１８	下午	名詞
２４	委员	名詞
２３	委员会	名詞
５７	会	名詞
５７	会	動詞
５８	在	動詞
５８	在	前置詞
５８	在	副詞
３７	布宜诺斯艾利斯	名詞
４３	讨论	動詞
５９	了	機能語
５０	这个	代名詞
６０	这	代名詞
６１	个	名詞（分類辞）
５３	问题	名詞

表９：単一文字単語および複数文字単語を有する単語リスト

【００６０】

複数文字単語を単語リストに追加し、次いで単一文字単語を単語リストに追加することは、単語リストを作成する１つの例示的な方法にすぎないことを理解されたい。代替手法では、最初に単一文字単語を突き止め、次いで単語リストに複数文字単語を追加することによって単語リストを得ることができる。最初に単一文字単語を突き止めることに関しては、この手法も上述した図９に示される手法と同様である。ただし、ステップ９０３および９０４は省略される。具体的には、ステップ９０２で、機能は、字句知識ベース中のそのエントリに基づいて、文字が単一文字単語を構成するかどうか判定する。文字が単一文字単語を構成する場合は、機能はステップ９０５に続き、文字を単語リストに追加し、そうでない場合は文字を単語リストに追加することなくステップ９０６に続く。機能は、文章中の各文字を処理して、ステップ９０１、９０２、９０５、および９０６を通ってループすることによって文字が単語であるかどうか判定する。

【００６１】

代替手法では、次いで、機能が文章を処理して複数文字単語を突き止め、そのような単語を単語リストに追加する。機能は、図６に関して上述した方法を使用することができる。しかし、文章が、他の単語を含むことができない複数文字単語を含む可能性があるため、すなわち、複数文字単語に関するＩｇｎｏｒｅＰａｒｔｓ属性がセット状態である場合、複数文字単語を構成する単一文字単語を単語リストから削除または除去することが有益で

ある。単語リストからこれらの単一文字単語を除去することが、パーサ１３３に要求される分析を最小限に抑える。

【００６２】

しかし、ＩｇｎｏｒｅＰａｒｔｓ属性がセット状態である２つの複数文字単語が重複している場合、単語リストからの単一文字単語の除去は複雑である。一般的な例が参考になる。文字シーケンスＡＢＣが考察下の文章中に存在し、ＩｇｎｏｒｅＰａｒｔｓ属性がセット状態である複数文字単語ＡＢおよびＢＣを文章が備えることができると仮定する。また、Ａ、Ｂ、およびＣが単一文字単語であると仮定する。単語ＡＢおよびＢＣによってカバーされる単一文字単語全てが単に単語リストから除去される場合、問題がある。具体的には、ＢＣが文章中の正しい単語である場合、単語Ａが外されてしまう。同様に、ＡＢが文章中の正しい単語である場合、単語Ｃが外されてしまう。どちらの場合にも、文章を介する「パス」に壊れていないものがなくなるため、文章が構文解析されない。これが起こるのを防止するため、複数文字単語中の全ての単一文字単語が、重複している部分によってカバーされる（１つまたは複数の）単語を除いて、ＩｇｎｏｒｅＰａｒｔｓ属性の値に関わらず保持される。上述した一般例では、単語ＡとＣはどちらも単語リスト中に保持される。しかし、Ｂは、シーケンスの重複部分であるため、単語リストから除去される。図８を参照すると、代替手法において機能がステップ８０２に達した場合、単語候補を構成する全ての単一文字単語がリストから除去されることになる。そうではなく、機能がステップ８０６に達した場合、重複していない単一文字単語が保持され、重複している（１つまたは複数の）部分が除去される。

【００６３】

上述の方法では、単語候補中の最後の文字に関するＮｅｘｔＣｈａｒリストを検査し（ステップ８０１）、その最後の文字がない状態での単語候補が単語であるかどうか確かめる（ステップ８０４）ことによって、起こり得る重複する単語が突き止められる。代替手法では、重複する単語を、単語リストと共にパーサ１３３に提供される他の情報を検査することによって見つけることができる。具体的には、単語リストに加えて、パーサ１３３が単語リスト中の各単語の位置情報を受信する。表３の例からは、各文字が、１～２２まで順に番号付けされる。この位置情報を使用して、単語リスト中の各単語ごとに、単語の開始位置および単語の終了位置が求められる。例として表９において識別される単語を参照すると、番号「６」によって示される単語は、開始文字位置１および終了文字位置２を有し、番号「１２」によって示される単語は、開始文字位置２および終了文字位置３を有する。単一文字単語は、開始文字位置が終了文字位置と等しい。このとき、終了文字位置と、文章中であり得る近傍の単語の開始文字位置とを検査することによって、重複する単語を簡単に確かめることができる。具体的には、文章中であり得る単語の終了文字位置が、文章中で次にあり得る文字の開始文字位置よりも大きいまたはそれと等しい場合、重複状態が存在する。

【００６４】

複数文字単語および単一文字単語を単語リストに追加し、それらの単語に関する字句記録を生成した後、機能は、字句記録に確率を割り当て、これは構文解析プロセスにおいて字句記録を介してアプリケーションを順序付けるためにパーサによって使用される。以下で論ずる図１０および１１は、字句記録に確率を割り当てるために機能によって使用される２つの代替手法を示す。

【００６５】

図１０は、第１の手法による、単語リスト中の単語から生成された字句記録に確率を割り当てるために機能によって行われることが好ましいステップを示す流れ図である。機能は最終的に、構文解析プロセス中の早いうちにパーサに字句記録を考察させる高い確率値、または構文解析プロセス中の後の方でパーサに字句記録を考察させる低い確率値に、各字句記録に関する確率を設定することが好ましい。ステップ１００１～１００５で、機能は、単語リスト中の各単語を通ってループする。ステップ１００２で、現行単語が単語リスト中のより大きな単語に含まれている場合は、機能はステップ１００４に続き、そうでな

い場合はステップ１００３に続く。ステップ１００３で、機能は、単語を表す字句記録に関する確率を高い確率値に設定する。ステップ１００３の後、機能はステップ１００５に続く。ステップ１００４では、単語を表す字句記録に関する確率を低い確率値に設定する。ステップ１００４の後、機能はステップ１００５に続く。ステップ１００５で、処理すべきさらなる単語が単語リスト中に残っている場合は、機能はステップ１００１に続き、単語リスト中の次の単語を処理し、そうでない場合はこれらのステップが終了する。

【００６６】

以下の表１０は、図１０に示されるステップに従って単語リスト中の各単語に割り当てられた確率値を示す。この確率を検討することによって、機能が、各文字を含む少なくとも１つの単語に高い確率値を割り当て、それにより各文字を含む少なくとも１つの字句記録が構文解析プロセス中の早いうちに考察されることがわかる。

【００６７】

【表１０】

番号	単語	品詞	確率値
６	昨天	名詞	高
５４	昨	形態素	低
５５	天	名詞	低
１２	天下	名詞	高
５６	下	名詞（ローカライザ）	低
１８	下午	名詞	高
２４	委员	名詞	低
２３	委员会	名詞	高
５７	会	名詞	低
５７	会	動詞	低
５８	在	動詞	高
５８	在	前置詞	高
５８	在	副詞	高
３７	布宜诺斯艾利斯	名詞	高
４３	讨论	動詞	高
５９	了	機能語	高
５０	这个	代名詞	高
６０	这	代名詞	低
６１	个	名詞（分類辞）	低
５３	问题	名詞	高

表１０：確率を有する単語リスト

【００６８】

図１１は、第２の手法による、単語リスト中の単語から生成された字句記録に確率を割り当てるために機能によって行われることが好ましいステップを示す流れ図である。ステップ１１０１で、機能は、単語リストを使用して、全体が単語リスト中の単語から構成された文章の全てのあり得る区分化を識別する。ステップ１１０２で、機能は、ステップ１１０１で識別された、最も少ない単語を含む１つまたは複数のあり得る区分化を選択する。あり得る区分化の２つ以上が最少数の単語を有する場合、機能は、そのようなあり得る区分化をそれぞれ選択する。以下の表１１は、表９に示される単語リストから生成された、最少数の単語（９）を有するあり得る区分化を示す。

【００６９】

【表１１】

昨天下午委员会在布宜诺斯艾利斯讨论了这个问题。

表１１

【００７０】

ステップ１１０３で、機能は、（１つまたは複数の）選択された区分化での単語の字句記録に関する確率を高い確率値に設定する。ステップ１１０４で、機能は、（１つまたは複数の）選択された区分化にはない単語の字句記録に関する確率を低い確率値に設定する。ステップ１１０４の後、これらのステップが終了する。

【００７１】

以下の表１２は、図１１に示されるステップに従って単語リスト中の各単語に割り当てられた確率値を示す。この確率を検討することによって、機能が、各文字を含む少なくとも１つの単語に高い確率値を割り当て、それにより各文字を含む少なくとも１つの字句記録が構文解析プロセス中の早いうちに考察されることがわかる。

【００７２】

【表１２】

番号	単語	品詞	確率値
６	昨天	名詞	高
５４	昨	形態素	低
５５	天	名詞	低
１２	天下	名詞	高
５６	下	名詞（ローカライザ）	低
１８	下午	名詞	高
２４	委员	名詞	低
２３	委员会	名詞	高
５７	会	名詞	低
５７	会	動詞	低
５８	在	動詞	高
５８	在	前置詞	高
５８	在	副詞	高
３７	布宜诺斯艾利斯	名詞	高
４３	讨论	動詞	高
５９	了	機能語	高
５０	这个	代名詞	高
６０	这	代名詞	低
６１	个	名詞（分類辞）	低
５３	问题	名詞	高

表１２：確率を有する単語リスト

【００７３】

本発明の１つの広範な態様では、確率を、重複する単語の対に割り当てることもできる。一般的な文字シーケンスＡＢＣでは、統計データが、単語ＡＢとＣの組合せの確率がＡとＢＣの組合せよりも高いことを示す場合がある。このとき、パーサ１３３は、まず組合せＡＢとＣを考察すべきであり、ＡとＢＣの組合せは、ＡＢとＣを使用して正常な分析を見つけることができないことがわかるまで考察されるべきでない。統計データはまた、あり得る組合せＡＢとＣ、またはＡとＢＣの１つが不可能であることを示すこともできる。

【００７４】

重複する単語の対にある単語に相対確率を割り当てるため、または不可能な組合せを除去

するために、字句知識ベース１３２に情報が格納される。特に、字句知識ベース１３２中の多くの複数文字単語に追加のリストを関連付けることができる。リストは以下のものを含む。

【００７５】

（１）第１の左条件リスト－このエントリにある単語は、文章中でこのリストにある文字の１つによってすぐ前に先行された場合に低い確率を割り当てられる。

【００７６】

（２）第１の右条件リスト－このエントリにある単語は、文章中でこのリスト中にある文字の１つによってすぐ後に後続された場合に低い確率を割り当てられる。

【００７７】

（３）第２の左条件リスト－このエントリにある単語は、文章中でこのリスト中にある文字の１つによってすぐ前に先行された場合に無視される。すなわち、単語リスト中の複数文字単語がこの条件を満たしている場合、単語リストから除去される。

【００７８】

（４）第２の右条件リスト－このエントリにある単語は、文章中でこのリスト中にある文字の１つによってすぐ後に後続された場合に無視される。すなわち、単語リスト中の複数文字単語がこの条件を満たしている場合、単語リストから除去される。

【００７９】

前述の各リストが字句知識ベース１３２中のあらゆる複数文字単語に関しては存在しない場合があることに留意されたい。すなわち、字句知識ベース１３２中のある複数文字単語が、前述のリストのどれも有さない場合があり、その一方である複数文字単語は、リストの１つ、いくつか、または全てを有する。希望するなら、他のリストを、直前のまたは直後の文字に基づいて生成することができる。例えば、高い確率を割り当てるようにリストを生成することができる。リストは、字句知識ベース１３２に手動で入力される。

【００８０】

上述したように明確さを解決するために字句知識ベースを使用する分析に加え、構文解析を始める前に字句分析と共に規則ベースの明確さ分析を使用することもできる。例えば、文章中に文字列ＡＢＣＤが存在し、ＡＢ、ＢＣ、およびＣＤが全てあり得る単語である場合、ＡＢが先行する単語と重ならず、ＣＤが後続の単語と重ならず、ＡまたはＤが単語でなく、ＡＢＣもＢＣＤも単語でない場合には、単語ＢＣを無視する（単語リストから除去する）ことができる。

【００８１】

しかし、区分化のあいまいさを解決するパーサの能力と、上述した字句明確さとの間に論理的な依存関係がないことを強調したい。字句レベルでの単語の排除は、構文解析の複雑さを低減するが、必ずしも文章の正常な分析に必要な条件であるとは限らない。文章中の全ての正しい単語が機能１３１によって提供される単語リスト中にあり、単語リスト中の単語の数がパーサ１３３に過剰な負担になるほど大きくない限り、構文解析は正常である。したがって、字句明確さの完全な成功は、正しい単語区分化を含めた文章分析の成功を大幅に容易にするが、後者は前者に依存しない。これは、構成要素間に対話があるにもかかわらず、機能１３１とパーサ１３３を独立して開発することを可能にする。

【００８２】

図１２は、サンプル文章の構文構造を表すパーサによって生成された構文木を示す構文木図である。構文木は、そのヘッドに単一文章記録１２３１を有し、そのリーフにいくつかの字句記録１２０１～１２１１を有する階層構造であることがわかる。構文木はさらに、それぞれ単語を表す字句記録を１つまたは複数の単語を表すより大きな構文構造に組み合わせる中間構文記録１２２１～１２２７を有する。例えば、前置詞句記録１２２３は、前置詞を表す字句記録１２０４と、名詞を表す字句記録１２０６とを組み合わせる。図５のステップ５０６に従って、機能は、サンプル文章が区分化されるべき単語として構文木で字句記録１２０１～１２１１によって表される単語を識別する。この構文木は、文章に対するさらなる自然言語処理を実施するために、機能によって保持することもできる。

【００８３】
本発明を好ましい実施形態を参照しながら示し、記述してきたが、本発明の範囲を逸脱することなく、形式および詳細に様々な変更または修正を行うことができることを当業者は理解されよう。例えば、機能の態様を、中国語以外の言語での単語区分化を行うために適用することができる。さらに、本明細書に記述した技法の適切なサブセットまたはスーパーセットを単語区分化を行うために適用することができる。

【図面の簡単な説明】

【図１】　機能が実行することが好ましい汎用コンピュータシステムの高レベルブロック図である。

【図２】　機能が動作することが好ましい２つの段階を示す概略流れ図である。

【図３】　単語分割を実行するために使用される情報を含むように初期化段階において語彙知識ベースを増大させるために機能によって実行されることが好ましいステップを示す流れ図である。

【図４】　特定の単語が他のより小さい単語を含有することができるかどうかを判定するために実行されることが好ましいステップを示す流れ図である。

【図５】　センテンスをその成分語に分割するために機能によって実行されることが好ましいステップの流れ図である。

【図６】　単語リストに複数文字語を追加するために機能によって実行されることが好ましいステップの流れ図である。

【図７】　単語候補のＮｅｘｔＣｈａｒ状態とＣｈａｒＰｏｓ状態をテストするために機能によって実行されることが好ましいステップの流れ図である。

【図８】　現在の単語候補の最後の文字が、単語であることがある別の単語候補と重なるかどうかを判定するために機能によって実行されることが好ましいステップの流れ図である。

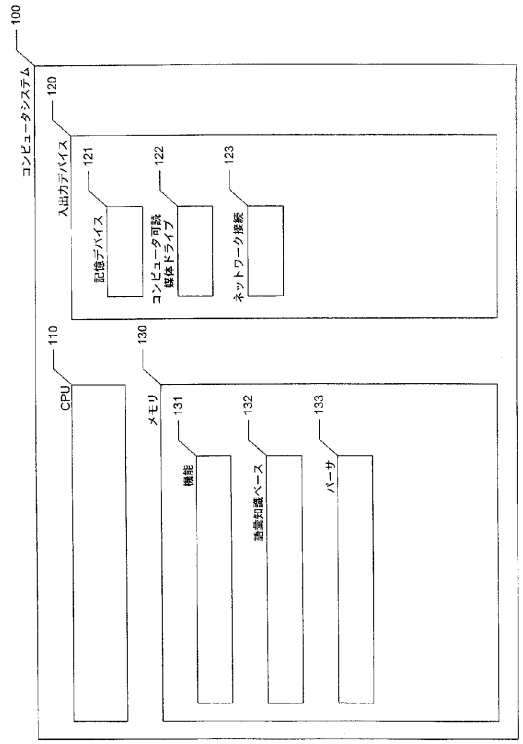
【図９】　単語リストに単数文字語を追加するために機能によって実行されることが好ましいステップの流れ図である。

【図１０】　第１の手法に従って単語リスト中の単語から生成された語彙記録に確率を割り振るために機能によって実行されることが好ましいステップの流れ図である。

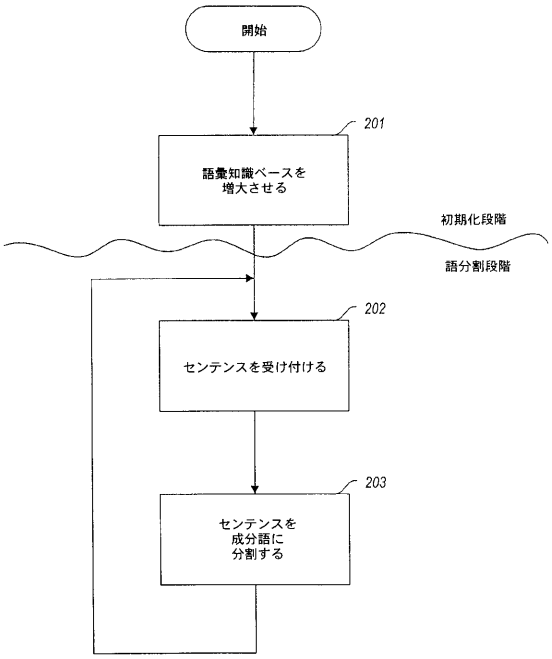
【図１１】　第２の手法に従って単語リスト中の単語から生成された語彙記録に確率を割り振るために機能によって実行されることが好ましいステップの流れ図である。

【図１２】　サンプルセンテンスの構文構造を表すパーサによって生成されたパースツリーを示すパースツリー図である。

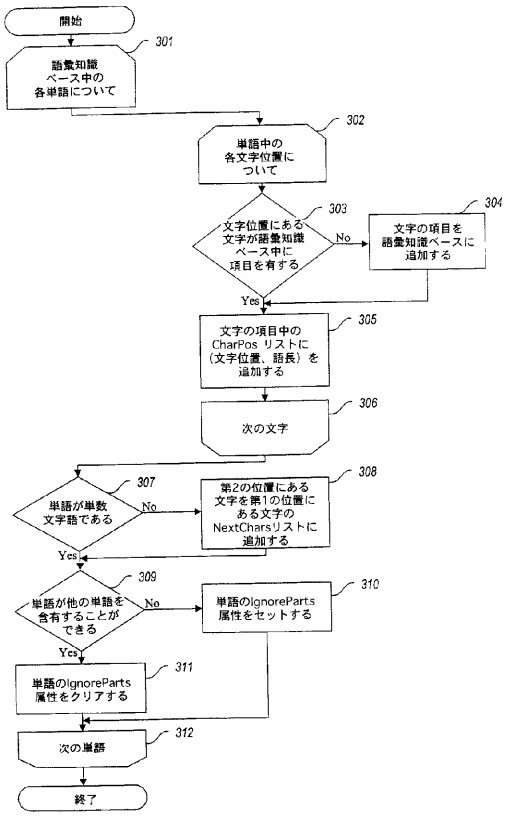
【図１】


100
コンピュータシステム
120
入出力デバイス
121
記憶デバイス
122
コンピュータ可読媒体ドライブ
123
ネットワーク接続
110
CPU
130
メモリ
131
機能
132
語彙知識ベース
133
パーサ

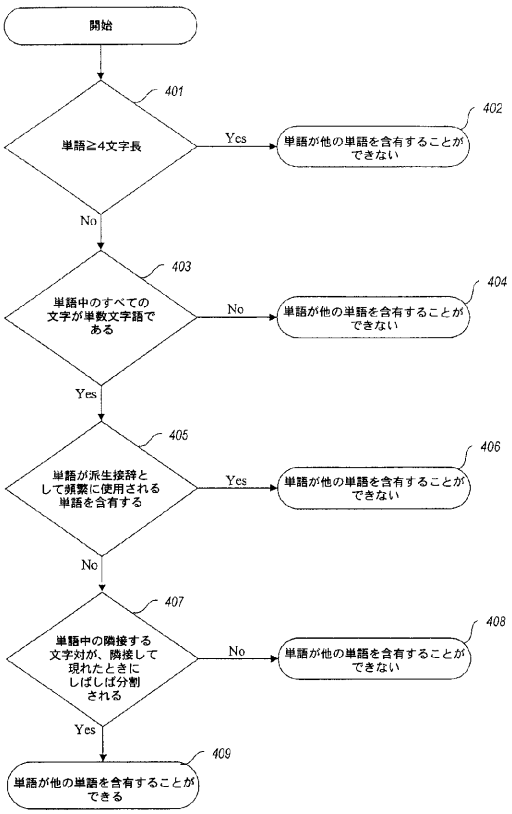
【図２】


開始
201
語彙知識ベースを増大させる
初期化段階
語分割段階
202
センテンスを受け付ける
203
センテンスを成分語に分割する

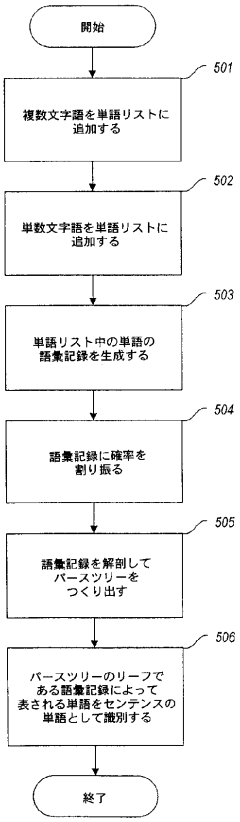
【図３】


開始
301
語彙知識ベース中の各単語について
302
単語中の各文字位置について
303
文字位置にある文字が語彙知識ベース中に項目を有する
No
304
文字の項目を語彙知識ベースに追加する
Yes
305
文字の項目中のCharPos リストに（文字位置、語長）を追加する
306
次の文字
307
単語が単数文字語である
No
308
第2の位置にある文字を第1の位置にある文字のNextCharsリストに追加する
Yes
309
単語が他の単語を含有することができる
No
310
単語のIgnoreParts属性をセットする
Yes
311
単語のIgnoreParts属性をクリアする
312
次の単語
終了

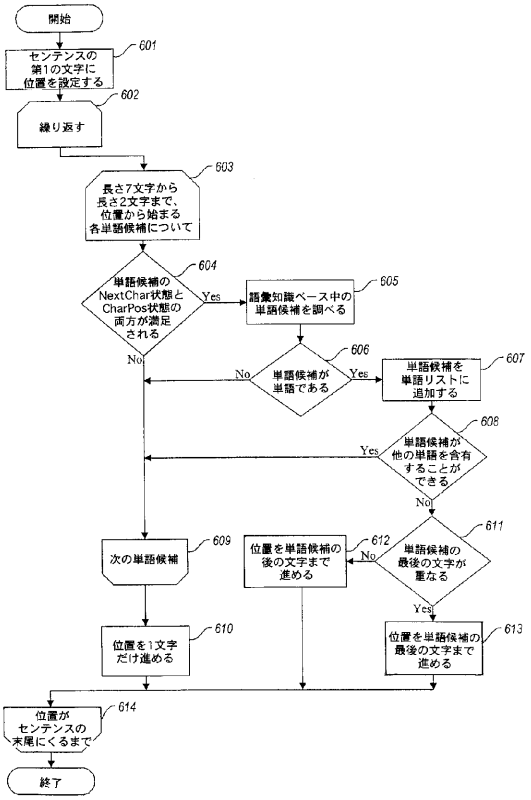
【図４】


開始
401
単語≧4文字長
Yes
402
単語が他の単語を含有することができない
No
403
単語中のすべての文字が単数文字語である
No
404
単語が他の単語を含有することができない
Yes
405
単語が派生接辞として頻繁に使用される単語を含有する
Yes
406
単語が他の単語を含有することができない
No
407
単語中の隣接する文字対が、隣接して現れたときにしばしば分割される
No
408
単語が他の単語を含有することができない
Yes
409
単語が他の単語を含有することができる

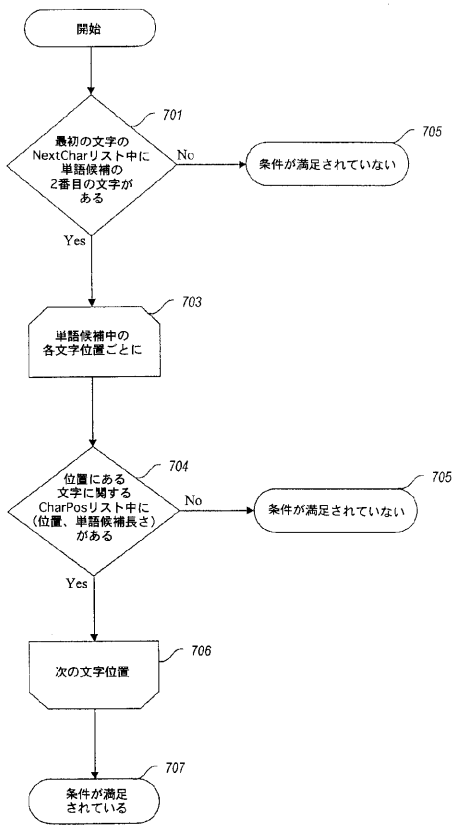
【図５】


開始
501
複数文字語を単語リストに追加する
502
単数文字語を単語リストに追加する
503
単語リスト中の単語の語彙記録を生成する
504
語彙記録に確率を割り振る
505
語彙記録を解剖してパースツリーをつくり出す
506
パースツリーのリーフである語彙記録によって表される単語をセンテンスの単語として識別する
終了

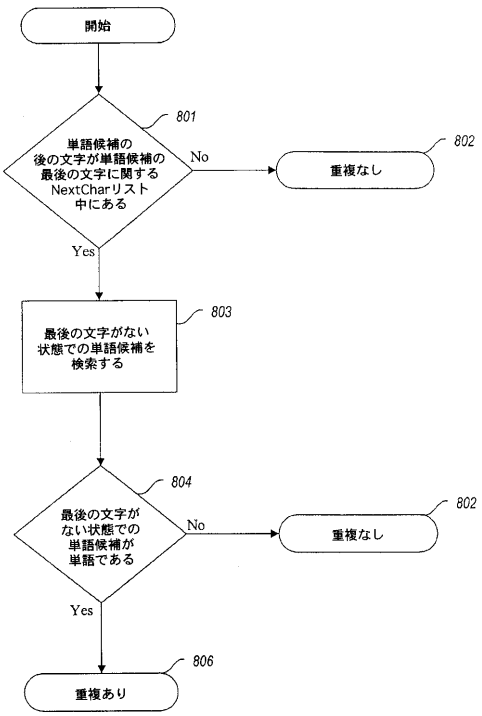
【図６】


開始
601
センテンスの第1の文字に位置を設定する
602
繰り返す
603
長さ7文字から長さ2文字まで、位置から始まる各単語候補について
604
単語候補のNextChar状態とCharPos状態の両方が満足される
Yes
No
605
語彙知識ベース中の単語候補を調べる
606
単語候補が単語である
No
Yes
607
単語候補を単語リストに追加する
608
単語候補が他の単語を含有することができる
Yes
No
609
次の単語候補
610
位置を1文字だけ進める
611
単語候補の最後の文字が重なる
612
位置を単語候補の後の文字まで進める
No
Yes
613
位置を単語候補の最後の文字まで進める
614
位置がセンテンスの末尾にくるまで
終了

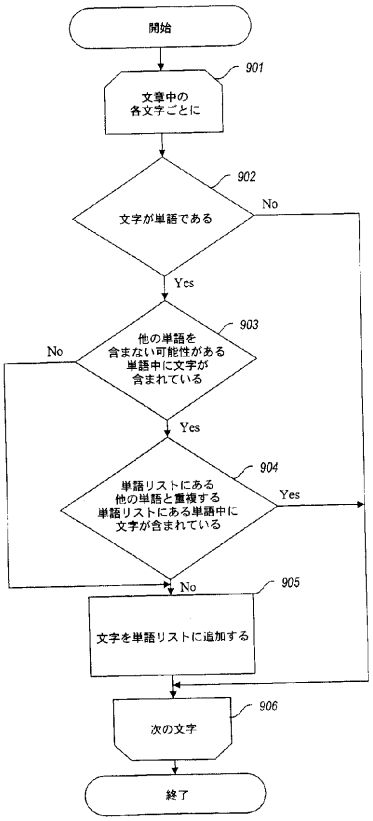
【図７】


開始
701
最初の文字のNextCharリスト中に単語候補の2番目の文字がある
No
Yes
705
条件が満足されていない
703
単語候補中の各文字位置ごとに
704
位置にある文字に関するCharPosリスト中に(位置、単語候補長さ)がある
No
Yes
705
条件が満足されていない
706
次の文字位置
707
条件が満足されている

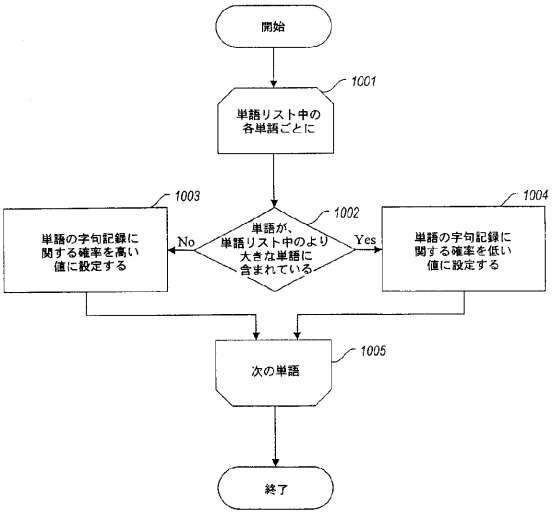
【図８】


開始
801
単語候補の後の文字が単語候補の最後の文字に関するNextCharリスト中にある
No
Yes
802
重複なし
803
最後の文字がない状態での単語候補を検索する
804
最後の文字がない状態での単語候補が単語である
No
Yes
802
重複なし
806
重複あり

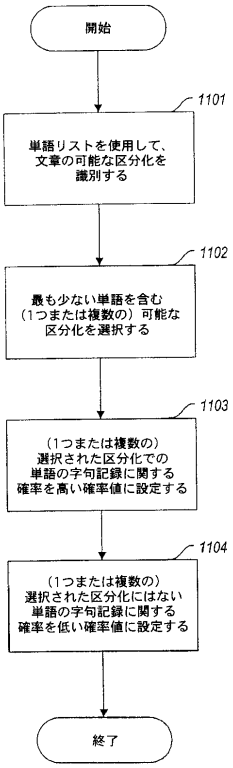
【図９】



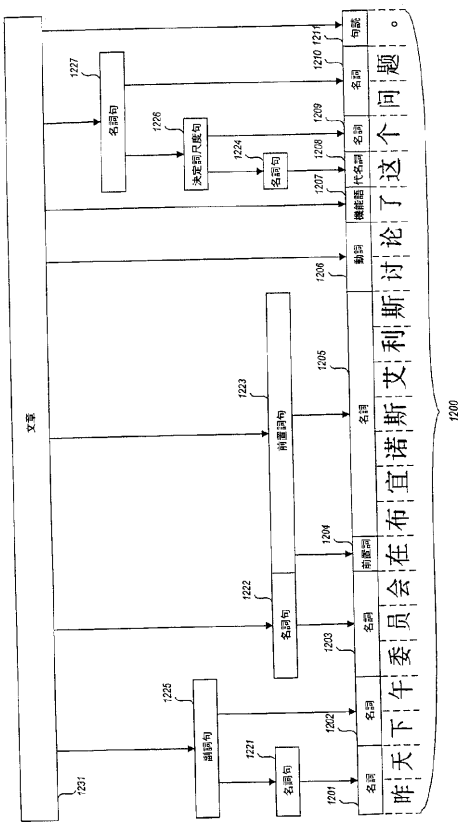
【図１０】



【図１１】



【図１２】



フロントページの続き

(72)発明者 リチャードソン スティーブン ディー.
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド ノースイースト １３２ ストリート １８０２８
(72)発明者 ジアン ジシン
アメリカ合衆国 ９８０５２ ワシントン州 レッドモンド １７６ プレイス ノースイースト ナンバー３ ９３０７

合議体
審判長 小曳 満昭
審判官 池田 聡史
審判官 飯田 清司

(56)参考文献 特開平６－３２５０７６（ＪＰ，Ａ）
特開平１－２７４２７２（ＪＰ，Ａ）
特開平８－３３９３８３（ＪＰ，Ａ）
Ｃｈｉｎｇ－Ｌｏｎｇ Ｙｅｈ ａｎｄ Ｈｓｉ－Ｊｉａｎ Ｌｅｅ、ＲＵＬＥ－ＢＡＳＥＤ ＷＯＲＤ ＩＤＥＮＴＩＦＩＣＡＴＩＯＮ ＦＯＲ ＭＡＮＤＡＲＩＮ ＣＨＩＮＥＳＥ ＳＥＮＴＥＮＣＥＳ － Ａ ＵＮＩＦＩＣＡＴＩＯＮ ＡＰＰＲＯＡＣＨ、ＣＯＭＰＵＴＥＲ ＰＲＯＣＥＳＳＩＮＧ ＯＦ ＣＨＩＮＥＳＥ ＆ ＯＲＩＥＮＴＡＬ ＬＡＮＧＵＡＧＥＳ、米国、１９９１．０３．発行、Ｖｏｌ．５，Ｎｏ．２、ｐ９７－１１８

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F17/21,17/27