(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2004-118461
(P2004-118461A)
(43)公開日　平成16年4月15日(2004.4.15)

(51)Int.Cl.[7]	F I			テーマコード（参考）
G06F 17/22	G06F	17/22	514U	5B009
G06F 17/21	G06F	17/21	550A	

審査請求 未請求　請求項の数 10　OL　（全 17 頁）

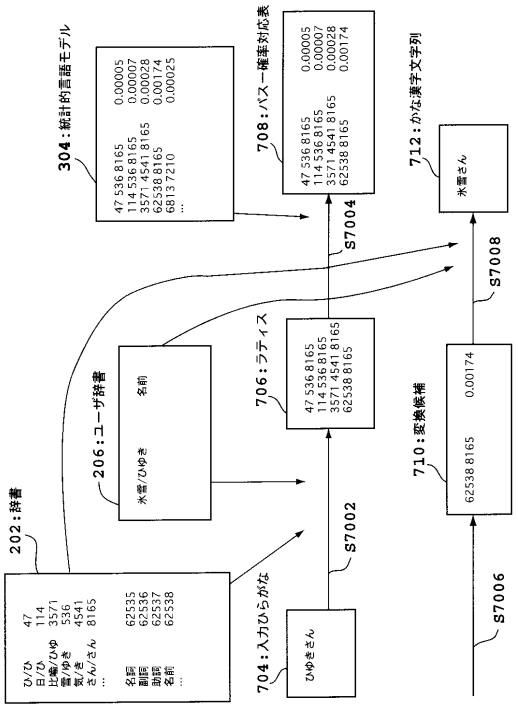
(21)出願番号　特願2002-279934(P2002-279934)
(22)出願日　平成14年9月25日(2002.9.25)

(71)出願人　391055933
マイクロソフト　コーポレイション
ＭＩＣＲＯＳＯＦＴ　ＣＯＲＰＯＲＡＴＩＯＮ
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　（番地なし）
(74)代理人　100077481
弁理士　谷　義一
(74)代理人　100088915
弁理士　阿部　和夫
(74)代理人　100106998
弁理士　橋本　傳一

最終頁に続く

(54)【発明の名称】言語モデルのトレーニング方法、かな漢字変換方法、言語モデルのトレーニング装置、かな漢字変換装置、コンピュータプログラムおよびコンピュータ読み取り可能な記録媒体

(57)【要約】
【課題】統計的言語モデルに基づきながら、品詞によって定義された語も変換できるようにしたかな漢字変換装置を提供する。
【解決手段】ステップＳ７００２において、コンピュータシステムは、辞書２０２とユーザ辞書とを用いて、入力ひらがな７０４から語ＩＤと品詞ＩＤの混ざったＩＤの組み合わせ（パス）の集合（ラティス）を作成する。ステップＳ７００４において、統計的言語モデル３０４から、各パスが生起する確率を取り出し、各パスと確率を対応付けたパス－確率対応表７０８を生成する。ステップＳ７００６では、パス－確率対応表７０８のうちから、最も確率の高いパスを変換候補７１０として選択する。そして、ステップＳ７００８では、辞書２０２とユーザ辞書２０６とを用いて、選択されたパスをかな漢字文字列７１２に変換する。
【選択図】　図７



【特許請求の範囲】
【請求項１】
文字列を使用したコンピュータによる言語モデルのトレーニング方法であって、
表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、前記文字列に含まれる語および品詞に識別子を付与するステップと、
前記文字列の中で、前記付与された識別子が特定の順序で生起する確率を示す言語モデルを生成するステップと
を備えることを特徴とする言語モデルのトレーニング方法。
【請求項２】
請求項１に記載の言語モデルのトレーニング方法により生成された言語モデルを用いたコンピュータによるかな漢字変換方法であって、
表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、入力された文字列から、識別子の組み合わせの集合を作成するステップと、
前記言語モデルから、前記作成された識別子の組み合わせの各々が生起する確率を取り出すステップと、
前記取り出された確率が最も高い識別子の組み合わせを選択するステップと、前記辞書を用いて、前記選択された識別子の組み合わせをかなまたは漢字に変換するステップと
を備えることを特徴とするかな漢字変換方法。
【請求項３】
請求項２に記載のかな漢字変換方法において、前記コンピュータは語と品詞との対応がユーザまたはベンダにより登録される登録辞書を有し、前記変換するステップは、前記辞書と前記登録辞書とを用いて前記識別子の組み合わせを変換することを特徴とするかな漢字変換方法。
【請求項４】
語および品詞を含む文字列を使用した言語モデルのトレーニング装置であって、
表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を記憶した記憶手段と、
前記記憶手段に記憶された辞書を用いて、前記文字列に含まれる語および品詞に識別子を付与する付与手段と、
前記文字列の中で、前記付与手段により付与された識別子が特定の順序で生起する確率を示す言語モデルを生成する生成手段と
を備えることを特徴とする言語モデルのトレーニング装置。
【請求項５】
請求項４に記載の言語モデルのトレーニング装置により生成された言語モデルを用いたかな漢字変換装置であって、
表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を記憶した記憶手段と、
前記記憶手段に記憶された辞書を用いて、入力された文字列から、識別子の組み合わせの集合を作成する作成手段と、
前記言語モデルから、前記作成手段により作成された識別子の組み合わせの各々が生起する確率を取り出す取出手段と、
前記取出手段により取り出された確率が最も高い識別子の組み合わせを選択する選択手段と、
前記辞書を用いて、前記選択手段により選択された識別子の組み合わせをかなまたは漢字に変換する変換手段と
を備えることを特徴とするかな漢字変換装置。
【請求項６】
請求項５に記載のかな漢字変換装置において、前記記憶手段は語と品詞との対応がユーザまたはベンダにより登録される登録辞書を更に記憶し、前記変換手段は、前記記憶手段に記憶された辞書と登録辞書とを用いて前記識別子の組み合わせを変換することを特徴とす

るかな漢字変換装置。
【請求項７】
語および品詞を含む文字列を使用した言語モデルのトレーニングを行うためのコンピュータプログラムであって、コンピュータに対し、
表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、前記文字列に含まれる語および品詞に識別子を付与するステップと、
前記文字列の中で、前記付与された識別子が特定の順序で生起する確率を示す言語モデルを生成するステップと
を実行させることを特徴とするコンピュータプログラム。
【請求項８】
請求項７に記載のコンピュータプログラムにより生成されたされた言語モデルを有するコンピュータに対し、
表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、入力された文字列から、識別子の組み合わせの集合を作成するステップと、
前記言語モデルから、前記作成された識別子の組み合わせの各々が生起する確率を取り出すステップと、
前記取り出された確率が最も高い識別子の組み合わせを選択するステップと、
前記辞書を用いて、前記選択された識別子の組み合わせをかなまたは漢字に変換するステップと
を実行させることを特徴とするコンピュータプログラム。
【請求項９】
請求項８に記載のコンピュータプログラムにおいて、前記コンピュータは語と品詞との対応がユーザまたはベンダにより登録される登録辞書を有し、前記変換するステップは、前記辞書と前記登録辞書とを用いて前記識別子の組み合わせを変換することを特徴とするコンピュータプログラム。
【請求項１０】
請求項７から９のいずれかに記載のコンピュータプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、言語モデルのトレーニング方法、かな漢字変換方法、言語モデルのトレーニング装置、かな漢字変換装置、コンピュータプログラムおよびコンピュータ読み取り可能な記録媒体に関し、より詳細には、語と品詞の混ざった文字列を使用して統計的言語モデルで品詞を扱う言語モデルのトレーニング方法、かな漢字変換方法、言語モデルのトレーニング装置、かな漢字変換装置、コンピュータプログラムおよびコンピュータ読み取り可能な記録媒体に関する。
【０００２】
【従来の技術】
日本語の文字列を入力する装置として、キーボードから入力したいかな漢字文字列に対応するかな文字列を入力し、漢字変換キーの入力に応答して、かな文字列をかな漢字文字列に変換するパーソナルコンピュータやワード・プロセッサなどのかな漢字変換装置が従来から知られている。この装置に入力したかな文字列をかな漢字文字列に変換する場合は、漢字変換用の特定の１つまたは複数のキーを組み合わせて押下し、かな漢字文字列の候補を表示する。また、連続して候補を表示することも可能であり、この場合前候補キーや次候補キーを押下するなどして、漢字を確定することができる。かな文字列をカタカナ文字列に変換する場合や、ローマ字文字列に変換する場合も、上記と同様の手順で行われる。
【０００３】
入力された文字列についてかな漢字変換を行うかな漢字変換装置では、文字列に対応するかな漢字の候補を決定するのに、形態素や各フレームの解析情報を参照することによって

、変換精度を高めている。形態素とは、１つ以上の音素からなる意味をもった最小の言語単位をいい、形態素解析では、文字列に含まれている形態素の切れ目を認識し、および形態素の品詞を認定する。

【０００４】

このような日本語の形態素解析において、従来から接続コスト最小法に基づく変換処理が知られている。これは、文の単語分割に対して何らかの接続コストを設定し、文全体で接続コストの和が最小となるような単語分割を選択する方法である。従って、表記、読み、品詞さえ指定されれば、どのような語も変換できる（例えば、非特許文献１参照）。

【０００５】

接続コスト最小法では、品詞接続コストと単語コストを定義する。品詞接続コストは接続がまれな品詞間ほど大きく、単語コストは出現頻度が小さいほど大きくなるように設定する。しかし、接続コスト最小法では、コストを設定するための方法論が存在しない。

【０００６】

この欠点を解消するための方法として、統計的言語モデルの研究が盛んに行われている（例えば、非特許文献１参照）。これは、接続コスト最小法のコストに相当する言語モデルを対象領域のテキストから自動的に学習する方法であり、情報理論と確率理論とに基づく明確な理論的根拠を備え、かつ実験的にも高い精度を持っている。

【０００７】

【非特許文献１】

田中穂積監修、「自然言語処理－基礎と応用－」電子情報通信学会、平成１１年３月２５日

【０００８】

【発明が解決しようとする課題】

統計的言語モデルに基づいたかな漢字変換システムでは、実世界において各語が生起する確率が指定されなければならない。しかしながら、各語が生起する確率を知りようがないユーザにとっては、辞書に対して確率を指定することは不可能である。このため、語に対して表記、読み、品詞が指定されていても、実世界における確率が指定されていなければ、変換処理を行うことができないという問題があった。

【０００９】

本発明はこのような問題に鑑みてなされたものであり、その目的とするところは、統計的言語モデルに基づきながら、品詞によって定義された語も変換できるかな漢字変換装置およびかな漢字変換方法を提供することにある。

【００１０】

【課題を解決するための手段】

このような目的を達成するために、請求項１に記載の発明は、文字列を使用したコンピュータによる言語モデルのトレーニング方法であって、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、前記文字列に含まれる語および品詞に識別子を付与するステップと、前記文字列の中で、前記付与された識別子が特定の順序で生起する確率を示す言語モデルを生成するステップとを備えることを特徴とする。

【００１１】

また、請求項２に記載の発明は、請求項１に記載の言語モデルのトレーニング方法により生成された言語モデルを用いたコンピュータによるかな漢字変換方法であって、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、入力された文字列から、識別子の組み合わせの集合を作成するステップと、前記言語モデルから、前記作成された識別子の組み合わせの各々が生起する確率を取り出すステップと、前記取り出された確率が最も高い識別子の組み合わせを選択するステップと、前記辞書を用いて、前記選択された識別子の組み合わせをかなまたは漢字に変換するステップとを備えることを特徴とする。

【００１２】

また、請求項３に記載の発明は、請求項２に記載のかな漢字変換方法において、前記コン

ピュータは語と品詞との対応がユーザまたはベンダにより登録される登録辞書を有し、前記変換するステップは、前記辞書と前記登録辞書とを用いて前記識別子の組み合わせを変換することを特徴とする。

【００１３】

また、請求項４に記載の発明は、語および品詞を含む文字列を使用した言語モデルのトレーニング装置であって、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を記憶した記憶手段と、前記記憶手段に記憶された辞書を用いて、前記文字列に含まれる語および品詞に識別子を付与する付与手段と、前記文字列の中で、前記付与手段により付与された識別子が特定の順序で生起する確率を示す言語モデルを生成する生成手段とを備えることを特徴とする。

【００１４】

また、請求項５に記載の発明は、請求項４に記載の言語モデルのトレーニング装置により生成された言語モデルを用いたかな漢字変換装置であって、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を記憶した記憶手段と、前記記憶手段に記憶された辞書を用いて、入力された文字列から、識別子の組み合わせの集合を作成する作成手段と、前記言語モデルから、前記作成手段により作成された識別子の組み合わせの各々が生起する確率を取り出す取出手段と、前記取出手段により取り出された確率が最も高い識別子の組み合わせを選択する選択手段と、前記辞書を用いて、前記選択手段により選択された識別子の組み合わせをかなまたは漢字に変換する変換手段とを備えることを特徴とする。

【００１５】

また、請求項６に記載の発明は、請求項５に記載のかな漢字変換装置において、前記記憶手段は語と品詞との対応がユーザまたはベンダにより登録される登録辞書を更に記憶し、前記変換手段は、前記記憶手段に記憶された辞書と登録辞書とを用いて前記識別子の組み合わせを変換することを特徴とする。

【００１６】

また、請求項７に記載の発明は、語および品詞を含む文字列を使用した言語モデルのトレーニングを行うためのコンピュータプログラムであって、コンピュータに対し、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、前記文字列に含まれる語および品詞に識別子を付与するステップと、前記文字列の中で、前記付与された識別子が特定の順序で生起する確率を示す言語モデルを生成するステップとを実行させることを特徴とする。

【００１７】

また、請求項８に記載の発明は、請求項７に記載のコンピュータプログラムにより生成されたされた言語モデルを有するコンピュータに対し、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、入力された文字列から、識別子の組み合わせの集合を作成するステップと、前記言語モデルから、前記作成された識別子の組み合わせの各々が生起する確率を取り出すステップと、前記取り出された確率が最も高い識別子の組み合わせを選択するステップと、前記辞書を用いて、前記選択された識別子の組み合わせをかなまたは漢字に変換するステップとを実行させることを特徴とする。

【００１８】

また、請求項９に記載の発明は、請求項８に記載のコンピュータプログラムにおいて、前記コンピュータは語と品詞との対応がユーザまたはベンダにより登録される登録辞書を有し、前記変換するステップは、前記辞書と前記登録辞書とを用いて前記識別子の組み合わせを変換することを特徴とする。

【００１９】

また、請求項１０に記載の発明は、コンピュータ読み取り可能な記録媒体であって、請求項７から９のいずれかに記載のコンピュータプログラムを記録したことを特徴とする。

【００２０】

【発明の実施の形態】

本発明の好ましい実施形態を示す以下の説明では、本明細書の一部を形成し、本発明を実践できる特定の実施形態を示す添付図面を参照する。本発明の範囲から逸脱せずに、他の実施形態を使用することができ、構造的変更を行うことができることを理解されたい。

【００２１】

図１に、本発明を実施することができる適切なコンピュータシステム１００の一例を示す。このコンピュータシステム１００は適切なコンピュータシステムの一例にすぎず、本発明の使用法または機能性の範囲に関するいかなる限定をも示唆しようとするものではない。また、コンピュータシステム１００は、同図に示す任意の１つまたは複数の構成要素の組み合わせに関する依存または要件を有するものと解釈されるべきではない。

【００２２】

本発明は、コンピュータで実行中の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コンテキストに関して説明することができる。一般に、プログラムモジュールは、特定タスクを実行するか、または特定の抽象データタイプを実施するルーチン、プログラム、オブジェクト、構成要素、データ構造などを含む。本発明は、通信ネットワークを介してリンクされている遠隔処理デバイスによってタスクが実施される分散型コンピュータシステムで実施することもできる。分散型コンピュータシステムでは、プログラムモジュールは、記憶装置を含めて、ローカルおよび遠隔コンピュータの記録媒体内に配置することができる。

【００２３】

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の形態による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の構成要素は、限定はしないが、処理ユニット１２０、システムメモリ１３０、およびシステムメモリを含む様々なシステム構成要素を処理ユニット１２０に結合するシステムバス１２１を含む。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む複数バス構造のうちのどれであっても良い。限定的ではないが、例示として、このようなアーキテクチャには、Ｉｎｄｕｓｔｒｙ Ｓｔｎｄａｒｄ Ａｒｃｈｉｔｅｃｔｕｒｅ（ＩＳＡ）バス、Ｍｉｃｒｏ Ｃｈａｎｎｅｌ Ａｒｃｈｉｔｅｃｔｕｒｅ（ＭＣＡ）バス、Ｅｎｈａｎｃｅｄ ＩＳＡ（ＥＩＳＡ）バス、Ｖｉｄｅｏ Ｅｌｅｃｔｒｏｎｉｃｓ Ｓｔａｎｄａｒｄｓ Ａｓｓｏｃｉａｔｅ（ＶＥＳＡ）ローカルバス、およびＭｅｚｚａｎｉｎｅバスとしても知られているＰｅｒｉｐｈｅｒａｌ Ｃｏｍｐｏｎｅｎｔ Ｉｎｔｅｒｃｏｎｎｅｃｔ（ＰＣＩ）バスが含まれる。

【００２４】

コンピュータ１１０は、一般に様々なコンピュータ読み取り可能な記録媒体を含む。コンピュータ読み取り可能な記録媒体は、コンピュータ１１０によってアクセス可能ないかなる使用可能な媒体であってもよく、揮発性媒体および不揮発性媒体、取り外し可能媒体および取り外し不可能媒体の両方を含むものである。限定はしないが例示として、コンピュータ読み取り可能な記録媒体は、コンピュータ読み取り可能な記録媒体と通信媒体とを含むことができる。

【００２５】

コンピュータ読み取り可能な記録媒体は、コンピュータ実行可能命令、データ構造、プログラムモジュールまたは他のデータなどの情報格納用のいかなる方法または技術で実施される揮発性および不揮発性、取り外し可能および取り外し不可能媒体の両方を含む。コンピュータ読み取り可能な記録媒体は、限定はしないが、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリまたは他のメモリ技術、ＣＤ－ＲＯＭ、デジタル多目的ディスク（ＤＶＤ）または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または他の磁気記憶装置、所望の情報を格納するために使用され、またコンピュータ１１０によってアクセスすることができる他のいかなる媒体をも含む。

【００２６】

通信媒体は、通常、コンピュータ実行可能命令、データ構造、プログラムモジュールまた

は搬送波または他の搬送メカニズムなどの変調されたデータ信号形式の他のデータを組み込むものであり、いかなる情報伝達媒体をも含むものである。「変調されたデータ信号」という用語は、１つまたは複数のその特徴的な設定を有する信号、または信号中に情報を符号化するような方法で変更された信号を意味している。限定はしないが例示として、通信媒体は、有線ネットワークまたは直接有線接続などの有線媒体と、音波、ＲＦ、赤外線または他の無線媒体などの無線媒体を含む。コンピュータ読み取り可能な記録媒体の範囲には、上記のいかなる組み合わせも含まれるべきである。

【００２７】

システムメモリ１３０は、読み出し専用メモリ（ＲＯＭ）１３１およびランダムアクセスメモリ（ＲＡＭ）１３２などの揮発性または不揮発性メモリ形式のコンピュータ読み取り可能な記録媒体を含む。起動中などに、コンピュータ１１０内の要素間で情報を転送するために役立つ基本ルーチンを含んでいる基本入出力システム１３３（ＢＩＯＳ）は、一般にＲＯＭ１３１内に格納されている。ＲＡＭ１３２は、一般に、処理ユニット１２０に即時アクセス可能な、またはその時点において処理ユニット１２０により操作されているデータまたはプログラムモジュールを含む。限定はしないが例示として、図１に、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を示す。

【００２８】

コンピュータ１１０は、他の取り外し可能／取り外し不可能、揮発性／不揮発性コンピュータ読み取り可能な記録媒体を含むこともできる。例示として、図１に、取り外し不可能な不揮発性磁気媒体に対して読み出しまたは書込みするハードディスクドライブ１４１、取り外し可能な不揮発性磁気ディスク１５２に対して読み出しまたは書込みする磁気ディスクドライブ１５１、取り外し可能な不揮発性光ディスク１５６に対して読み出しまたは書込みをする、ＣＤ－ＲＯＭまたは他の光学媒体などの光ディスクドライブ１５５を示す。例示的オペレーティング環境で使用することができる他の取り外し可能／取り外し不可能な揮発性／不揮発性コンピュータ読み取り可能な記録媒体は、限定はしないが、磁気テープカセット、フラッシュメモリカード、デジタル多目的ディスク、デジタルビデオテープ、半導体ＲＡＭ、半導体ＲＯＭなどを含む。ハードディスクドライブ１４１は、一般に、インターフェース１４０などの取り外し不可能なメモリインターフェースを介してシステムバス１２１に接続されており、磁気ディスクドライブ１５１と光ディスクドライブ１５５は、一般に、インターフェース１５０などの取り外し可能なメモリインターフェースによってシステムバス１２１に接続されている。

【００２９】

図１に示す上記のドライブおよびそれらに関連したコンピュータ読み取り可能な記録媒体は、コンピュータ実行可能命令、データ構造、プログラムモジュールおよびコンピュータ１１０のための他のデータの記憶装置を提供する。図１では、例えば、ハードディスクドライブ１４１は、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。これらの構成要素は、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と同じであっても異なっていても良いという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、その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、最低限、それらが異なるコピーであることを示すために異なる番号が与えられている。ユーザは、キーボード１６２および、一般にマウス、トラックボールまたはタッチパッドと呼ばれるポインティングデバイス１６１などの入力デバイスによってコンピュータ１１０にコマンドおよび情報を入力することができる。他の入力デバイス（図示せず）は、マイクロフォン、ジョイスティック、ゲームパッド、衛星放送アンテナ、スキャナなどを含むことができる。これらのおよび他の入力デバイスは、システムバスに結合されたユーザ入力インターフェース１６０を介して処理ユニット１２０に接続されることがしばしばあるが、パラレルポート、ゲームポートまたはユニ

バーサルシリアルバス（ＵＳＢ）などの他のインターフェースおよびバス構造に接続されることもできる。モニタ１９１または他のタイプの表示装置も、ビデオインターフェース１９０などのインターフェースを介してシステムバス１２１に接続される。モニタの他に、コンピュータは、出力周辺インターフェース１９５を介して接続することができるスピーカ１９７およびプリンタ１９６など、他の周辺出力装置を含むこともできる。

【００３０】

コンピュータ１１０は、遠隔コンピュータ１８０などの１つまたは複数の遠隔コンピュータへの論理接続を使用してネットワーク化された環境において動作することができる。遠隔コンピュータ１８０は、別のパーソナルコンピュータ、サーバ、ルータ、ネットワークＰＣ、ピアデバイスまたは他の共通ネットワークノードであってよく、図１では記憶装置１８１しか示していないが、一般にコンピュータ１１０に関して上記で説明した多くのまたはすべての要素を含む。図１で示す論理接続は、ローカルエリアネットワーク（ＬＡＮ）１７１およびワイドエリアネットワーク（ＷＡＮ）１７３を含むが、他のネットワークを含むこともできる。このようなネットワーキング環境は、事務所、企業全体に巡らされているコンピュータネットワーク、イントラネットおよびインターネットでは一般的なことである。

【００３１】

ＬＡＮネットワーキング環境で使用されるとき、コンピュータ１１０は、ネットワークインターフェースまたはアダプタ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とき、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モデム１７２または他の手段を含む。内部であっても外部であっても良いモデム１７２は、ユーザ入力インターフェース１６０または他の適切なメカニズムを介してシステムバス１２１に接続することができる。ネットワーク環境において、コンピュータ１１０に関して示されたプログラムモジュールまたはその一部は、遠隔記憶装置に格納することができる。限定はしないが例示として、図１に、記憶装置１８１上に常駐するものとして遠隔アプリケーションプログラム１８５を示す。図示するネットワーク接続は一例であり、コンピュータ間で通信リンクを確立する他の手段を使用することもできることを理解されたい。

【００３２】

以下の説明では、本発明は、特に指摘しない限り、アプリケーションプログラムのコンピュータ実行可能命令をシステムメモリ１３１にロードした処理ユニット１２０が、そのコンピュータ実行可能命令に基づき実行することができる動作を説明する。この動作において、処理ユニット１２０はコンピュータ実行可能命令に基づきプログラムデータ１３７を参照し、あるいはその更新を行う。

【００３３】

従って、コンピュータによって実行されるときに時折表現されることのあるこのような動作および演算には、コンピュータの処理ユニットによる、構造化形式のデータを表現する電子信号の操作が含まれることを理解されたい。この操作はデータを変換するか、またはコンピュータのメモリシステム中の記憶場所にデータを維持し、そこで、当業者が良く理解している方法でコンピュータの演算を再構成あるいは変更する。データが維持されているデータ構造は、データ形式によって規定される特定の特性を有するメモリの物理的な記憶場所である。本発明を上記の条件で説明してはいるが、以下で説明する様々な動作および演算はハードウェアでも実施可能であることを当業者なら理解するように、この説明は限定を意図するものではない。

【００３４】

図２は、本実施形態に係るプログラムデータ１３７の内容をより詳細に示す図であり、本発明に関わる部分のみを概略的に示している。

【００３５】

プログラムデータ１３７は、コーパス２０２と、辞書２０４と、ユーザ辞書２０６とを含んでいる。コーパス２０２は、自然言語処理等に利用される大規模テキストデータであっ

て、文字列が形態素ごとに分割され、各形態素について品詞が決定された（即ち、品詞タグ付けされた）ものである。その他、係り受けなどの統語情報が付加されたものもコーパス２０２として利用することができる。辞書２０４は、語および品詞の各々に対する識別子（ＩＤ）を定義したデータである。ここで、語は文字の表記とその読みとを含んでいる。ユーザ辞書２０６は、ユーザ個人が使い勝手を良くするために単語や定型句を登録して作成する登録辞書の１つである。ここで、登録辞書は、ユーザ辞書の他、専門辞書や分野別辞書などのベンダにより登録されるものであっても良い。

【００３６】

図３は、本実施形態に係るコンピュータシステムによる言語モデルのトレーニング方法の概要を示す図である。以下では説明を簡単にするためにバイグラムモデルを例に挙げて説明するが、本発明はトライグラムモデル等の他のマルチグラムモデルにも適用可能であることはいうまでもない。

【００３７】

まず、コーパス２０２に含まれている文字列から、表記と読みとを含む語のＩＤ、および、品詞のＩＤを定義した辞書２０４を用いることにより、特定のＩＤペアの生起回数を示すＩＤペア－生起回数対応表３０２を生成する（Ｓ３００２）。このＩＤペアの生起回数から、各ＩＤペアがコーパス２０２内で生起する確率を示す統計的言語モデル３０４が生成される（Ｓ３００４）。以下、図４～図６を参照し、上述した方法の各ステップについて詳細に説明する。

【００３８】

図４に示すように、辞書２０４は、語データ４０２と、品詞データ４０４とを有する。語データ４０２は、「は／は」、「だ／だ」、「今日／きょう」、のように、表記と読みとを含む語４０６の集合である語データ４０２と、ＩＤ４０８との対応を示すものである。また、語データ４０２には、文頭および文末とＩＤ４０８との対応も定義されている。品詞データ４０４は、名詞、副詞、助詞といった品詞４１０の各々のＩＤ４１２を示すものである。

【００３９】

いま、図５に示すようにコーパス２０２が

文頭　今日／きょう／名詞　は／は／助詞　天気／てんき／副詞　だ／だ／助動詞　。／。／句点　文末

という情報を含んでいるものとする。ステップＳ３００２において、コンピュータシステムは辞書２０４を使用し、コーパス２０２に含まれる文字列「今日は天気だ。」に含まれる語および品詞に、辞書２０４内のＩＤを付与する。次いで、コーパス２０２内で、特定のＩＤペアが生起する回数５１０を数える。ここで、特定のＩＤペアとは、語ＩＤと語ＩＤのペア５０２、品詞ＩＤと品詞ＩＤのペア５０４、語ＩＤと品詞ＩＤのペア５０６、および品詞ＩＤと語ＩＤのペア５０８である。

【００４０】

次いで、ステップＳ３００４では、図６に示すように、ＩＤペア－生起回数対応表３０２内の各ＩＤペアの生起回数５１０に基づき、特定のＩＤのペアが生起する確率６０２を示す統計的言語モデル３０４を生成する。

【００４１】

以上説明したように、本実施形態では、コーパス内の語と品詞とが混ざった少なくとも１つの文字列から、語と品詞のペアが生起する確率、品詞同士のペアが生起する確率、および語同士のペアが生起する確率を計算する。このように言語モデルをトレーニングすることにより、語の確率情報による変換とともに品詞情報を用いた変換も可能となる。以下、上述したように生成された言語モデルを用いたかな漢字変換方法について説明する。

【００４２】

図７は、本実施形態に係るかな漢字変換方法の概要を示す図である。なお、以下の説明において使用されるコンピュータシステムとして図１に示すものが使用されるが、これは上述した言語モデルのトレーニング方法に使用されるものと同一であっても良く、異なるも

のであっても良い。後者の場合、上述のように生成された言語モデルは、ＣＤ－ＲＯＭ等の取り外し可能不揮発性メモリ、あるいは有線ネットワーク等の通信媒体を使用して、以下の処理を実行するコンピュータシステムへ提供することができる。

【００４３】

ステップＳ７００２において、コンピュータシステムは、辞書２０２とユーザ辞書とを用いて、入力ひらがな７０４から語ＩＤと品詞ＩＤの混ざったＩＤの組み合わせ（パス）の集合（ラティス）を作成する。ステップＳ７００４において、統計的言語モデル３０４から、各パスの生起する確率を取り出し、各パスと確率を対応付けたパス－確率対応表７０８を生成する。ステップＳ７００６では、パス－確率対応表７０８のうちから、最も確率の高いパスを変換候補７１０として選択する。そして、ステップＳ７００８では、辞書２０２とユーザ辞書２０６とを用いて、選択されたパスをかな漢字文字列７１２に変換する。以下、図８～１０を参照し、本実施形態に係るかな漢字変換方法について詳細に説明する。

【００４４】

まず、図８に示すように、キーボード１６２等の入力手段を介して、品詞と語とを含む入力ひらがな７０４がコンピュータシステムに入力される。入力ひらがな７０４は、入力と同時にモニタ１９１に表示される。ステップＳ７００２では、入力ひらがな７０４から各種の形態素が生成され、辞書２０２およびユーザ辞書２０６からあらゆる可能性の変換すべき語の候補が取り出され、語ＩＤまたは品詞ＩＤの組み合わせ（パス）の集合であるラティス７０６を生成する。

【００４５】

例えば、１文字目「ひ」の候補は「ひ／ひ／４７」と「日／ひ／１１４」の２つである。また、これに続く候補は「雪／ゆき／５３６」である。最後の２文字の候補は「さん／さん／８１６５」である。ここまでで、

４７　５３６　８１６５

１１４　５３６　８１６５

というパスができる。また、同時にユーザ辞書やベンダ辞書も参照され、ＩＤで候補が作成され、ラティスに追加される。同図に示す例では、ユーザ辞書２０６中に、「ひゆき」という読みに対して「氷雪／名前」という表記および品詞が対応付けられている。辞書２０２によれば「名前」という品詞のＩＤは６２５３８なので、このＩＤと「さん／さん／８１６５」という語のＩＤと組み合わせた

６２５３８　８１６５

というパスがラティス７０６に追加される。なお、辞書２０２およびユーザ辞書２０６を参照する順序は逆であっても良い。

【００４６】

ステップＳ７００４では、図９に示すように、統計的言語モデル３０４からラティス７０６に含まれるパスの生起する確率が取り出される。例えば、最初のパスは

４７　５３６　８１６５

なので、その確率０．０００００５が取り出される。同様にして、ラティス７０６を構成する各パスの確率が取り出され、パス－確率対応表７０８が生成される。

【００４７】

ステップＳ７００６では、図１０に示すように、パス－確率対応表７０８のうちから、最も確率が高いパス

６２５３８　８１６５

を変換候補７１０として選択する。

【００４８】

そして、図１１に示すように、ステップＳ７００８でまず変換候補７１０の最初のＩＤである６２５３８に対応する語または品詞を辞書２０２から取り出す。辞書２０２を参照すると、６２５３８は「名前」という品詞であることが分かる。本実施形態において、「名前」はユーザ辞書２０６に登録される品詞であることから、次にユーザ辞書２０６が参照

され、６２５３８というＩＤは、入力文字列「ひゆき」に対応する表記「氷雪」に変換される。

【００４９】

次のＩＤは８１６５なので、このＩＤに対応する語または品詞が辞書２０２から取り出される。ここで、８１６５というＩＤは「さん」という表記に変換される。このようにして、最終的に「氷雪さん」という文字列が得られ、モニタ１９１にかな漢字文字列７１２が表示される。

【００５０】

以上、本発明の好適な実施の形態について説明したが、本発明の前述の説明は、例示および説明を目的として提示されたものである。網羅的であること、または本発明を開示された正確な形態に制限することは、意図されていない。多数の修正形態および変形形態が、上の教示に鑑みて可能である。本発明の範囲は、この詳細な説明によるのではなく、請求項によって制限されることが意図されている。

【００５１】

【発明の効果】

以上説明したように、本発明では、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、文字列に含まれる語および品詞に識別子を付与するステップと、文字列の中で、付与された識別子が特定の順序で生起する確率を示す言語モデルを生成するステップとを備える。また、表記と読みとを含む語の識別子、および、品詞の識別子を定義した辞書を用いて、入力された文字列から、識別子の組み合わせの集合を作成するステップと、言語モデルから、作成された識別子の組み合わせの各々が生起する確率を取り出すステップと、取り出された確率が最も高い識別子の組み合わせを選択するステップと、辞書を用いて、選択された識別子の組み合わせをかなまたは漢字に変換するステップとを備える。従って、語の確率情報による変換とともに品詞情報を用いた変換も可能となる。

【００５２】

これは、特にサイズの小さい統計的言語モデルに有効である。統計的言語モデルのサイズを小さくするということは、即ち変換精度を落とすことを意味する。そこで、本発明を適用することにより、基本語ではない語を品詞ＩＤを使用して変換できるので、変換精度を高く保つことができる。

【００５３】

また、コンピュータは語と品詞との対応がユーザまたはベンダにより登録される登録辞書を有し、変換するステップは、辞書と登録辞書とを用いて識別子の組み合わせを変換するので、統計的言語モデルにおいて語ＩＤだけでなく品詞ＩＤもラティスに追加することで、ユーザ辞書やベンダ辞書に含まれている語の変換が可能となるという効果を奏する。

【図面の簡単な説明】

【図１】本発明を実施する例示的システムを構成するコンピュータシステムを示す図である。

【図２】本発明実施形態によるプログラムデータの内容を示すブロック図である。

【図３】本発明実施形態による言語モデルのトレーニング方法の動作の概要を示す図である。

【図４】本発明実施形態による言語モデルのトレーニング方法の動作を示す図である。

【図５】本発明実施形態による言語モデルのトレーニング方法の動作を示す図である。

【図６】本発明実施形態による言語モデルのトレーニング方法の動作を示す図である。

【図７】本発明実施形態によるかな漢字変換方法の動作の概要を示す図である。

【図８】本発明実施形態によるかな漢字変換方法の動作を示す図である。

【図９】本発明実施形態によるかな漢字変換方法の動作を示す図である。

【図１０】本発明実施形態によるかな漢字変換方法の動作を示す図である。

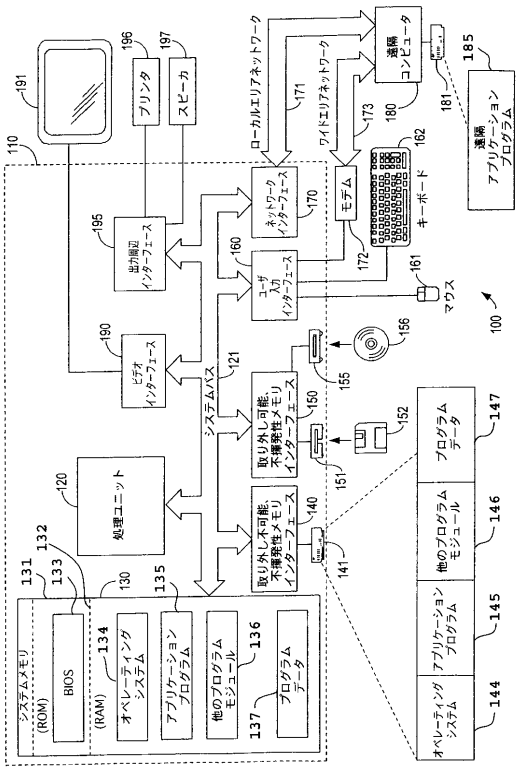
【図１１】本発明実施形態によるかな漢字変換方法の動作を示す図である。

【符号の説明】

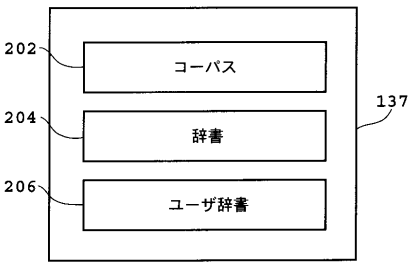
１００　コンピュータシステム
１１０　コンピュータ
１２０　処理ユニット
１２１　システムバス
１３０　システムメモリ
１３１　読み出し専用メモリ
１３２　ランダムアクセスメモリ
１３３　基本入出力システム
１３４　オペレーティングモジュール
１３５　アプリケーションプログラム
１３６　他のプログラムモジュール
１３７　プログラムデータ
１４０　取り外し不可能不揮発性メモリインターフェース
１４１　ハードディスクドライブ
１４４　オペレーティングシステム
１４５　アプリケーションプログラム
１４６　他のプログラムモジュール
１４７　プログラムデータ
１５０　取り外し可能不揮発性メモリインターフェース
１５１　磁気ディスクドライブ
１５２　取り外し可能な不揮発性磁気ディスク
１５５　光ディスクドライブ
１５６　取り外し可能な不揮発性光ディスク
１６０　ユーザ入力インターフェース
１６１　ポインティングデバイス
１６２　キーボード
１７０　アダプタ
１７１　ローカルエリアネットワーク（ＬＡＮ）
１７２　モデム
１７３　ワイドエリアネットワーク（ＷＡＮ）
１８０　遠隔コンピュータ
１８１　記憶装置
１８４　マルチレベルキャッシュ
１８５　遠隔アプリケーションプログラム
１９０　ビデオインターフェース
１９１　モニタ
１９５　出力周辺インターフェース
１９６　プリンタ
１９７　スピーカ
２０２　コーパス
２０４　辞書
２０６　ユーザ辞書
３０２　ＩＤペアー生起回数対応表
３０４　統計的言語モデル
４０２　語データ
４０４　品詞データ
４０６　表記／読み
４０８　ＩＤ番号
４１０　品詞
４１２　ＩＤ番号

５０２　語ＩＤ－語ＩＤのペア
５０４　品詞ＩＤ－品詞ＩＤのペア
５０６　語ＩＤ－品詞ＩＤのペア
５０８　品詞ＩＤ－語ＩＤのペア
５１０　ＩＤペアの生起回数
６０２　ＩＤペアの確率
７０４　入力ひらがな
７０６　ラティス
７０８　パス－確率対応表
７１０　変換候補
７１２　かな漢字文字列

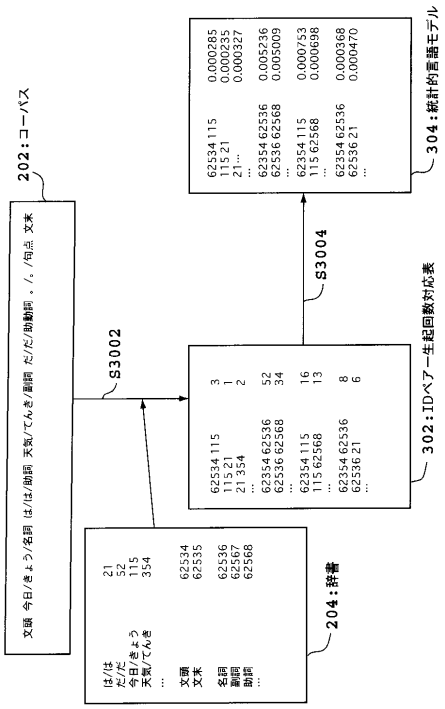
【図１】



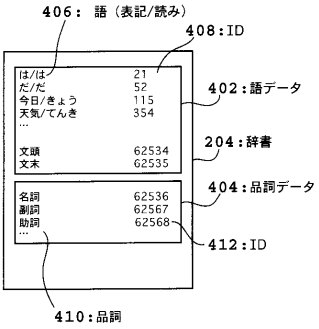
【図２】



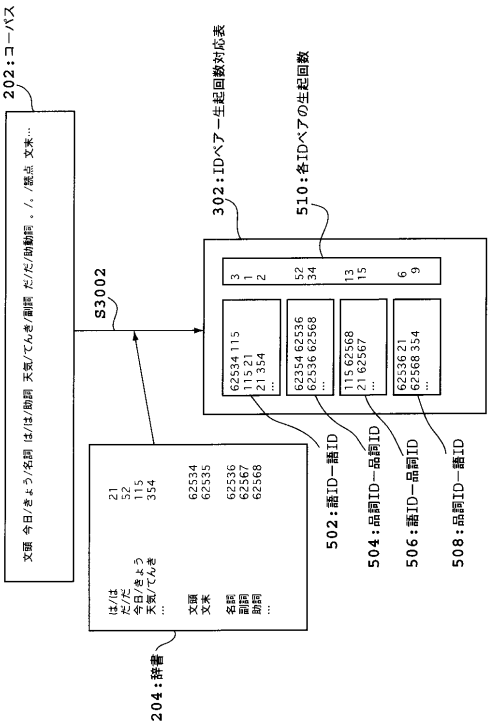
【図３】


202:コーパス
文頭 今日/きょう/名詞 は/は/助詞 天気/てんき/副詞 だ/だ/助動詞 。/。/句点 文末
S3002
S3004
204:辞書
302:IDペアー生起回数対応表
304:統計的言語モデル

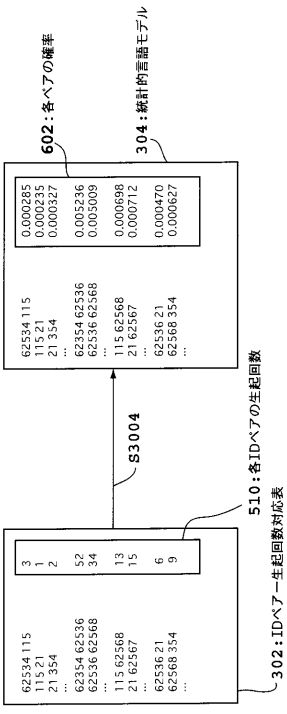
【図４】


406: 語（表記/読み）
408:ID
402:語データ
204:辞書
404:品詞データ
412:ID
410:品詞

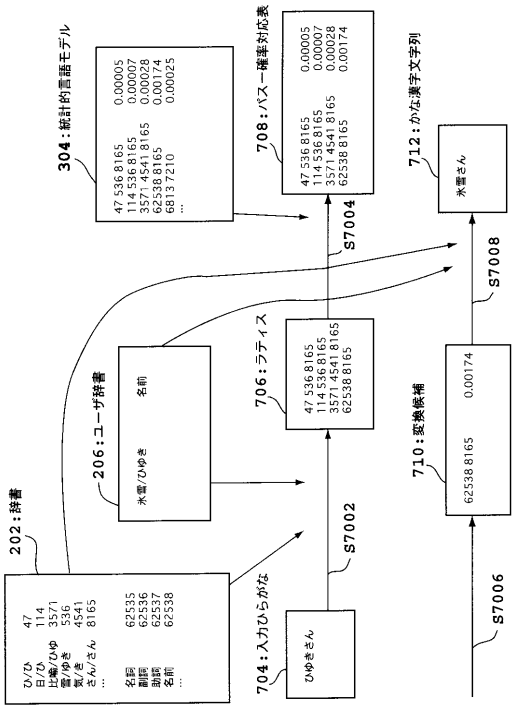
【図５】


202:コーパス
文頭 今日/きょう/名詞 は/は/助詞 天気/てんき/副詞 だ/だ/助動詞 。/。/感点 文末…
S3002
302:IDペアー生起回数対応表
510:各IDペアの生起回数
502:語ID−語ID
504:品詞ID−品詞ID
506:語ID−品詞ID
508:品詞ID−語ID
204:辞書

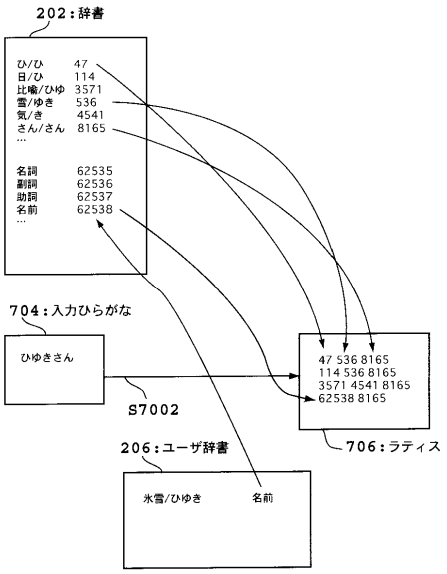
【図６】


602:各ペアの確率
304:統計的言語モデル
S3004
510:各IDペアの生起回数対応表
302:IDペアー生起回数対応表

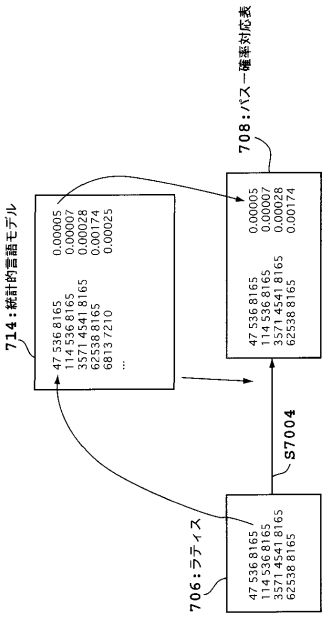
【図７】


202：辞書
ひ/ひ 47
日/ひ 114
比喩/ひゆ 3571
雪/ゆき 536
気/き 4541
さん/さん 8165
…
名詞 62535
副詞 62536
助詞 62537
名前 62538
…
206：ユーザ辞書
氷雪/ひゆき 名前
304：統計的言語モデル
47 536 8165 0.00005
114 536 8165 0.00007
3571 4541 8165 0.00028
62538 8165 0.00174
6813 7210 0.00025
…
708：パス－確率対応表
47 536 8165 0.00005
114 536 8165 0.00007
3571 4541 8165 0.00028
62538 8165 0.00174
706：ラティス
47 536 8165
114 536 8165
3571 4541 8165
62538 8165
704：入力ひらがな
ひゆきさん
710：変換候補
62538 8165 0.00174
712：かな漢字文字列
氷雪さん
S7002
S7004
S7006
S7008

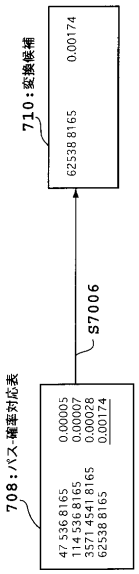
【図８】


202：辞書
ひ/ひ 47
日/ひ 114
比喩/ひゆ 3571
雪/ゆき 536
気/き 4541
さん/さん 8165
…
名詞 62535
副詞 62536
助詞 62537
名前 62538
…
704：入力ひらがな
ひゆきさん
S7002
47 536 8165
114 536 8165
3571 4541 8165
62538 8165
706：ラティス
206：ユーザ辞書
氷雪/ひゆき 名前

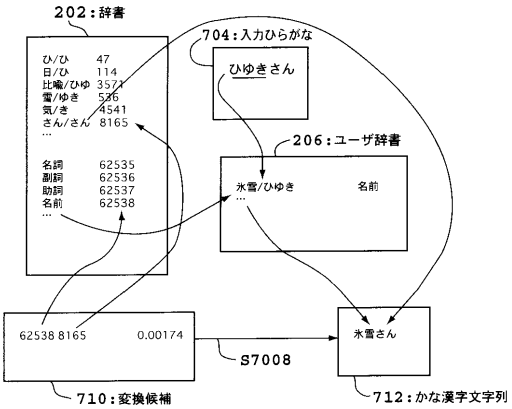
【図９】


714：統計的言語モデル
47 536 8165 0.00005
114 536 8165 0.00007
3571 4541 8165 0.00028
62538 8165 0.00174
6813 7210 0.00025
…
708：パス－確率対応表
47 536 8165 0.00005
114 536 8165 0.00007
3571 4541 8165 0.00028
62538 8165 0.00174
706：ラティス
47 536 8165
114 536 8165
3571 4541 8165
62538 8165
S7004

【図１０】


710：変換候補
62538 8165 0.00174
S7006
708：パス－確率対応表
47 536 8165 0.00005
114 536 8165 0.00007
3571 4541 8165 0.00028
62538 8165 0.00174

【図１１】


202：辞書
ひ/ひ 47
日/ひ 114
比喩/ひゆ 3571
雪/ゆき 526
気/き 4541
さん/さん 8165
…
名詞 62535
副詞 62536
助詞 62537
名前 62538
…
704：入力ひらがな
ひゆきさん
206：ユーザ辞書
氷雪/ひゆき 名前
…
62538 8165 0.00174
S7008
氷雪さん
710：変換候補
712：かな漢字文字列

フロントページの続き

(72)発明者　石橋　紀子
東京都調布市調布ケ丘１－１８－１　マイクロソフト株式会社　マイクロソフト調布技術センター内
(72)発明者　鹿子木　宏明
東京都調布市調布ケ丘１－１８－１　マイクロソフト株式会社　マイクロソフト調布技術センター内

Ｆターム(参考) 5B009 MA05　QA01　TA09