(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2005-92682
(P2005-92682A)
(43)公開日　平成17年4月7日(2005.4.7)

(51)Int.Cl.[7]	F I		テーマコード（参考）
G06F 17/28	G06F 17/28	P	5B091

審査請求 未請求　請求項の数 3　ＯＬ　（全 15 頁）

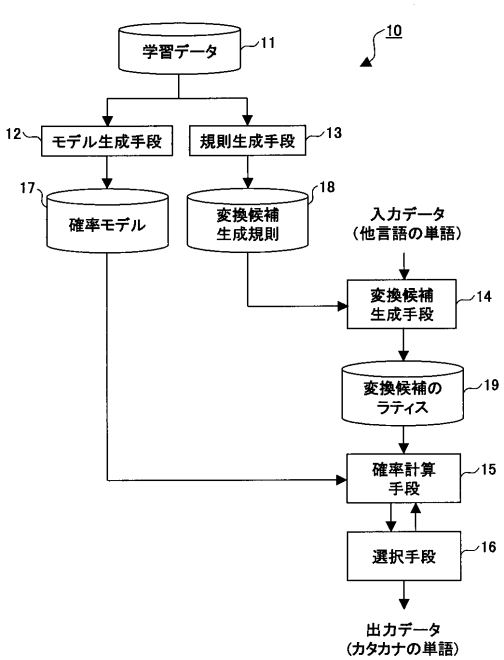
(21)出願番号　特願2003-327491 (P2003-327491)
(22)出願日　平成15年9月19日 (2003.9.19)

(71)出願人　000004352
日本放送協会
東京都渋谷区神南２丁目２番１号
(74)代理人　100070150
弁理士　伊東　忠彦
(72)発明者　後藤　功雄
東京都世田谷区砧一丁目１０番１１号　日本放送協会　放送技術研究所内
(72)発明者　加藤　直人
東京都世田谷区砧一丁目１０番１１号　日本放送協会　放送技術研究所内

Ｆターム(参考)　5B091 AB04　CA21　CC03　CC16　EA01

(54)【発明の名称】翻字装置、及び翻字プログラム

(57)【要約】
【課題】　他言語の単語からカタカナの単語への高精度な翻字を実現する。
【解決手段】　他言語の単語とカタカナの単語とにおける部分文字列が対応付けられたデータに基づいて、変換候補の規則を生成する変換候補規則生成手段と、前記他言語の単語を文脈情報に基づいて変換単位に分割するための分割確率を取得するモデルと、前記他言語とカタカナとの部分文字列の対応確率を計算するモデルとを生成するモデル生成手段と、前記他言語で入力される単語を前記変換候補規則生成手段により得られる変換規則に基づいて、前記カタカナの変換候補と前記他言語での変換単位とを生成する変換候補生成手段と、前記モデル生成手段により得られるモデルと、前記他言語と前記カタカナとの文脈情報とに基づいて、変換候補の生起確率を計算する確率計算手段と、前記確率計算手段により得られる生起確率が最大となる変換候補を選択する変換候補選択手段とを有する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
他言語の単語からカタカナの単語を生成するための翻字装置において、
前記他言語の単語と前記カタカナの単語とにおける部分文字列が対応付けられたデータに基づいて、変換候補の規則を生成する変換候補規則生成手段と、
前記他言語の単語を単語内の文脈情報に基づいて変換単位に分割するための分割確率を取得するモデルと、前記他言語とカタカナとの部分文字列の対応確率を単語内の文脈情報に基づいて計算するモデルとを生成するモデル生成手段と、
前記他言語で入力される単語を前記変換候補規則生成手段により得られる変換規則に基づいて、前記カタカナの変換候補と前記他言語での変換単位とを生成する変換候補生成手段と、
前記モデル生成手段により得られるモデルと、前記他言語と前記カタカナとの文脈情報とに基づいて、変換候補の生起確率を計算する確率計算手段と、
前記確率計算手段により得られる生起確率が最大となる変換候補を選択する変換候補選択手段とを有することを特徴とする翻字装置。
【請求項２】
前記変換候補選択手段は、
予め設定される評価式に基づいて、前記変換候補生成手段にて得られる変換候補からカタカナの変換候補を選択することを特徴とする請求項１に記載の翻字装置。
【請求項３】
他言語の単語からカタカナの単語を生成するための処理をコンピュータに実行させるための翻字プログラムにおいて、
前記他言語の単語と前記カタカナの単語とにおける部分文字列が対応付けられたデータに基づいて、変換候補の規則を生成する変換候補規則生成処理と、
前記他言語の単語を単語内の文脈情報に基づいて変換単位に分割するための分割確率を取得するモデルと、前記他言語とカタカナとの部分文字列の対応確率を単語内の文脈情報に基づいて計算するモデルとを生成するモデル生成処理と、
前記他言語で入力される単語を前記変換候補規則生成処理により得られる変換規則に基づいて、前記カタカナの変換候補と前記他言語での変換単位とを生成する変換候補生成処理と、
前記モデル生成処理により得られるモデルと、前記他言語と前記カタカナとの文脈情報とに基づいて、変換候補の生起確率を計算する確率計算処理と、
前記確率計算処理により得られる生起確率が最大となる変換候補を選択する変換候補選択処理とをコンピュータに実行させるための翻字プログラム。

【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、翻字装置、及び翻字プログラムに係り、特に、他言語の単語からカタカナの単語へ変換するための翻字装置、及び翻字プログラムに関する。
【背景技術】
【０００２】
従来、英語や日本語等の文字が異なる言語間において、固有名詞は多くの場合に元の単語の発音を表す外来語に翻訳される。特に、日本語では、カタカナを用いた単語に翻訳される場合が多い。
【０００３】
ここで、他言語からカタカナへの翻字処理に関する技術は、すでに開示されており、例えば、他言語を発音記号（音韻体系）へ変換してから、発音記号をカタカナへ変換する方式（例えば、非特許文献１参照。）や他言語からカタカナへ直接変換する小規模な変換テーブルを用いて変換する方式（例えば、非特許文献２参照。）がある。

【０００４】
また、カタカナ以外の文字への翻字も提案されている。例えば、「発音を考慮した変換単位」を用いる英語から韓国語への翻字処理の手法がある（例えば、非特許文献３参照）。また、英語から韓国語への翻字において、決定木を用いて変換単位の曖昧性を解消する手法がある（例えば、非特許文献４参照。）。更に、英語から韓国語への翻字において、決定木を用いて英語の文脈を考慮して変換候補の部分文字列を決定する手法がある（例えば、非特許文献５参照。）。
【非特許文献１】堀内　雄一，山崎　一生．１９９０．英単語のアルファベット表記から仮名表記への変換．情報処理学会自然言語処理研究会報告，Ｎｏ．７９－１，ｐｐ．１－８．
【非特許文献２】住吉　英樹，相沢　輝昭．英語固有名詞の片カナ変換．１９９４．情報処理学会論文誌，Ｖｏｌ．３５，　Ｎｏ．１，　ｐｐ．３５－４５．
【非特許文献３】Ｂｙｕｎｇ－Ｊｕ　Ｋａｎｇ　ａｎｄ　Ｋｅｙ－Ｓｕｎ　Ｃｈｏｉ．２０００．　Ａｕｔｏｍａｔｉｃ　Ｔｒａｎｓｌｉｔｅｒａｔｉｏｎ　ａｎｄ　Ｂａｃｋ－Ｔｒａｎｓｌｉｔｅｒａｔｉｏｎ　ｂｙ　Ｄｅｃｉｓｉｏｎ　Ｔｒｅｅ　Ｌｅａｒｎｉｎｇ．　Ｉｎｔｅｒｎａｔｉｏｎａｌ　Ｃｏｎｆｅｒｅｎｃｅ　ｏｎ　Ｌａｎｇｕａｇｅ　Ｒｅｓｏｕｒｃｅｓ　ａｎｄ　Ｅｖａｌｕａｔｉｏｎ，　ｐｐ．１１３５－１４１１．
【非特許文献４】Ｉｎ－Ｈｏ　Ｋａｎｇ　ａｎｄ　ＧｉｌＣｈａｎｇ　Ｋｉｍ．　２０００．　Ｅｎｇｌｉｓｈ－ｔｏ－Ｋｏｒｅａｎ　Ｔｒａｎｓｌｉｔｅｒａｔｉｏｎ　ｕｓｉｎｇ　Ｍｕｌｔｉｐｌｅ　Ｕｎｂｏｕｎｄｅｄ　Ｏｖｅｒｌａｐｐｉｎｇ　Ｐｈｏｎｅｍｅ　Ｃｈｕｎｋｓ．　Ｔｈｅ　１８ｔｈ　Ｉｎｔｅｒｎａｔｉｏｎａｌ　Ｃｏｎｆｅｒｅｎｃｅ　ｏｎ　Ｃｏｍｐｕｔａｔｉｏｎａｌ　Ｌｉｎｇｕｉｓｔｉｃｓ，　Ｖｏｌ．１，　ｐｐ．４１８－４２４．
【非特許文献５】Ｊｏｎｇ－Ｈｏｏｎ　Ｏｈ　ａｎｄ　Ｋｅｙ－Ｓｕｎ　Ｃｈｏｉ．　２００２．　Ａｎ　Ｅｎｇｌｉｓｈ－Ｋｏｒｅａｎ　Ｔｒａｎｓｌｉｔｅｒａｔｉｏｎ　Ｍｏｄｅｌ　ｕｓｉｎｇ　Ｐｒｏｎｕｎｃｉａｔｉｏｎ　ａｎｄ　Ｃｏｎｔｅｘｔｕａｌ　ｒｕｌｅｓ．　Ｔｈｅ　１９ｔｈ　Ｉｎｔｅｒｎａｔｉｏｎａｌ　Ｃｏｎｆｅｒｅｎｃｅ　ｏｎ　Ｃｏｍｐｕｔａｔｉｏｎａｌ　Ｌｉｎｇｕｉｓｔｉｃｓ．
【発明の開示】
【発明が解決しようとする課題】
【０００５】
しかしながら、非特許文献１に記載された技術は、他言語を発音記号へ変換することが困難であり、また非特許文献２に記載された技術は、小規模な変換テーブルを用いた方式の場合に詳細な文脈利用ができないため精度に問題があり、高精度な翻字処理を行うことはできない。
【０００６】
また、非特許文献３に記載された技術は、変換単位に複数の長さの部分文字列を用いる場合における変換元の部分文字列の選択の曖昧性を考慮しているが、更に高精度な変換を行うためには、翻字元と翻字先との文脈情報を考慮する必要がある。
【０００７】
更に、非特許文献４に記載された技術は、変換先の候補との対応関係を考慮せずに元の英語の単語の情報のみで一意に決定しているため、高精度な変換を行っているとはいえない。また、非特許文献５に記載された技術は、英語の１文字を変換の単位をした「発音を考慮しない変換単位」を用いているため、「発音を考慮した変換単位」と比べると精度が低下する。
【０００８】
本発明は、上述した問題点に鑑みなされたものであり、他言語からカタカナへの高精度な翻字を行うための翻字装置、及び翻字プログラムを提供することを目的とする。
【課題を解決するための手段】
【０００９】

上記課題を解決するために、本件発明は、以下の特徴を有する課題を解決するための手段を採用している。

【００１０】

請求項１に記載された発明は、他言語の単語からカタカナの単語を生成するための翻字装置において、前記他言語の単語と前記カタカナの単語とにおける部分文字列が対応付けられたデータに基づいて、変換候補の規則を生成する変換候補規則生成手段と、前記他言語の単語を単語内の文脈情報に基づいて変換単位に分割するための分割確率を取得するモデルと、前記他言語とカタカナとの部分文字列の対応確率を単語内の文脈情報に基づいて計算するモデルとを生成するモデル生成手段と、前記他言語で入力される単語を前記変換候補規則生成手段により得られる変換規則に基づいて、前記カタカナの変換候補と前記他言語での変換単位とを生成する変換候補生成手段と、前記モデル生成手段により得られるモデルと、前記他言語と前記カタカナとの文脈情報とに基づいて、変換候補の生起確率を計算する確率計算手段と、前記確率計算手段により得られる生起確率が最大となる変換候補を選択する変換候補選択手段とを有することを特徴とする。

【００１１】

請求項１記載の発明によれば、部分文字列が対応付けられたデータを利用して、変換規則を適用する単位となる変換単位への分割確率に基づいてカタカナの変換候補を選択することにより、他言語からカタカナへの翻字を高精度に行うことができる。また、モデル生成手段により得られるモデルと前記他言語と前記カタカナの文字との文脈情報とにより生起確率を算出することで、より高精度にカタカナの変換候補の選択を高精度に行うことができ、他言語からカタカナへの翻字を高精度に行うことができる。

【００１２】

請求項２に記載された発明は、前記変換候補選択手段は、予め設定される評価式に基づいて、前記変換候補生成手段にて得られる変換候補からカタカナの変換候補を選択することを特徴とする。

【００１３】

請求項２記載の発明によれば、予め設定される評価式を用いることで、一定の評価基準により容易に変換候補を選択することができる。これにより、他言語からカタカナへの翻字を高精度に行うことができる。

【００１４】

請求項３に記載された発明は、他言語の単語からカタカナの単語を生成するための処理をコンピュータに実行させるための翻字プログラムにおいて、前記他言語の単語と前記カタカナの単語とにおける部分文字列が対応付けられたデータに基づいて、変換候補の規則を生成する変換候補規則生成処理と、前記他言語の単語を単語内の文脈情報に基づいて変換単位に分割するための分割確率を取得するモデルと、前記他言語とカタカナとの部分文字列の対応確率を単語内の文脈情報に基づいて計算するモデルとを生成するモデル生成処理と、前記他言語で入力される単語を前記変換候補規則生成処理により得られる変換規則に基づいて、前記カタカナの変換候補と前記他言語での変換単位とを生成する変換候補生成処理と、前記モデル生成処理により得られるモデルと、前記他言語と前記カタカナとの文脈情報とに基づいて、変換候補の生起確率を計算する確率計算処理と、前記確率計算処理により得られる生起確率が最大となる変換候補を選択する変換候補選択処理とをコンピュータに実行させる。

【００１５】

請求項３記載の発明によれば、部分文字列が対応付けられたデータを利用して、変換規則を適用する単位となる変換単位への分割確率に基づいてカタカナの変換候補を選択することにより、他言語からカタカナへの翻字を高精度に行うことができる。また、モデル生成処理により得られるモデルと前記他言語と前記カタカナの文字との文脈情報とにより生起確率を算出することで、より高精度にカタカナの変換候補の選択を高精度に行うことができ、他言語からカタカナへの翻字を高精度に行うことができる。また、実行プログラムをコンピュータにインストールすることにより、容易に他言語からカタカナへの翻字を実

現することができる。
【発明の効果】
【００１６】
本発明によれば、他言語の単語からカタカナの単語への高精度な翻字を実現する。
【発明を実施するための最良の形態】
【００１７】
＜本発明の概要＞
本発明は、外来語のカタカナで表現される単語を、元の外国語（以後、他言語という）から生成するものである。そのために、カタカナの単語とその対訳の他言語の単語との両方の単語内で発音的に類似している部分を対応付けたデータベースに基づいて変換候補を生成し、その変換候補の適用スコアを統計的に学習し、学習結果を利用して翻字（音訳）を行う。
【００１８】
以下に、上記のような特徴を有する本発明における翻字装置、及び翻字プログラムを好適に実施した形態について、図面を用いて詳細に説明する。なお、本実施例では、他言語の例として、英語の場合について説明するが、韓国語等の言語においても本発明を適用することができる。
【００１９】
＜機能構成図＞
図１は、本発明における翻字装置の機能構成の一例を示す図である。図１の翻字装置１０は、学習データ１１と、モデル生成手段１２と、規則生成手段１３と、変換候補生成手段１４と、変換候補の確率計算手段１５と、最適な変換候補を選択する選択手段１６とを有するよう構成されている。
【００２０】
まず、モデル生成手段１２は、学習データ（コーパス）１１を入力して確率モデル１７を出力する。ここでは、最大エントロピー法に基づく学習を例として扱う。つまり、最大エントロピー法で利用する素性関数を定義して確率モデル１７を作成して出力する。なお、素性関数の定義内容については後述する。
【００２１】
規則生成手段１３は、単語内の部分文字列の対応がついた学習データ１１を用いて、英語の各部分文字列から変換されているカタカナの部分の字列の変換候補の規則を生成する。
【００２２】
例えば、「シ／ソー／ラ／ス：ｔｈｅ／ｓａｕ／ｒｕ／ｓ」のデータからは、“ｔｈｅ”→“シ”，“ｓａｕ”→“ソー”，“ｒｕ”→“ラ”，“ｓ”→“ス”という変換候補生成規則を得る。このような変換候補を学習データ１１中の全ての英語とカタカナの部分文字列の対応付けされたデータから生成し、変換候補生成規則１８を作成する。
【００２３】
なお、上述したモデル生成手段１２及び規則生成手段１３により確率モデル１７及び変換候補生成規則１８を作成するまでが学習フェーズとなる。つまり、学習データ１１に基づいて、入力される他言語の単語を翻字する処理の前に実行される。また、以下に説明する変換候補生成手段１４、確率計算手段１５、及び選択手段１６が、翻字を実行する実行フェーズとなる。
【００２４】
次に、実際の翻字を行う際には、例えば、他言語として英語の単語からなる入力データが変換候補生成手段１４に入力される。変換候補生成手段１４は、英語の単語からカタカナの単語を直接推定する。ここで、変換候補生成手段１４における候補生成の内容について具体的に説明する。
【００２５】
翻字を行う英語の単語からカタカナの単語の先頭に“＾”、単語の末尾に“＄”等の識

別子を追加して、英語の単語Ｅを以下に示す（１）式のように表現する。
【００２６】
【数１】

$$E = e_0^{m+1} = e_0 e_1 .. e_{m+1} \quad \cdots(1)$$

$$e_0 = \wedge \quad e_{m+1} = \$ \quad \cdots(2)$$

ここで、ｅ$_j$は、英語の単語のｊ番目の文字であり、ｍは、英語の単語の“＾”と“＄”以外の文字数である。また、ｅ$_0^{m+1}$は、ｅ$_0$からｅ$_{m+1}$までの文字列であることを示している。
【００２７】
この英語の単語の各部分に対する対応付けされた英語の部分文字列ｅｕ（Ｅｎｇｌｉｓｈ　Ｕｎｉｔ）と、カタカナの部分文字列ｋｕ（Ｋａｔａｋａｎａ　ｕｎｉｔ）とからなる変換候補生成規則の適用方法は、Ｅの文字列中に一致する変換候補生成規則のｅｕを全て適用し、そのｅｕに対応する全てのｋｕにより、ラティスＬ｛Ｋ｝を作成する。
【００２８】
ここで、一例として図２に英語の単語「ａｃｔｉｎｉｕｍ」の変換候補のラティスＬ｛Ｋ｝の例を示す。Ｌ｛Ｋ｝中の“＾”から“＄”までの各経路Ｐ$_d$∈（Ｐ$_1$，Ｐ$_2$，・・・，Ｐ$_q$）中の部分文字列を繋いだ文字列が変換先の単語の候補となる。例えば、図２において、「ｃ」には、「キ（ｋｉ）」、「ク（ｋｕ）」、及び「ック（ｋｋｕ）」の３つの候補があることを示している。なお、ｑは、Ｌ｛Ｋ｝中の“＾”から“＄”までの経路数を示している。
【００２９】
ここで、Ｌ｛Ｋ｝中のある経路Ｐ$_d$を選択した場合について説明する。この場合のＰ$_d$中の“＾”及び“＄”以外の部分文字列の数をｎ（Ｐ$_d$）とする。また、Ｐ$_d$中の部分文字列に、先頭から順番に番号を付与する。上述の条件により、Ｐ$_d$に対する英語の単語Ｅとその変換結果のカタカナの単語Ｋは、次のようになる。
【００３０】
【数２】

$$E = e_0^{m+1} = e_0 e_1 .. e_{m+1} = eu_0^{n(p_d)+1} = eu_0 eu_1 .. eu_{n(p_d)+1} \quad \cdots(3)$$

$$K = k_0^{l(p_d)+1} = k_0 k_1 .. k_{l(p_d)+1} = ku_0^{n(p_d)+1} = ku_0 ku_1 .. ku_{n(p_d)+1} \quad \cdots(4)$$

$$e_0 = k_0 = eu_0 = ku_0 = \wedge \ , \ e_{m+1} = k_{l(p_d)+1} = eu_{m+1} = ku_{n(p_d)} = \$ \quad \cdots(5)$$

ここで、ｋ$_j$はカタカナの単語のｊ番目の文字であり、ｍ（Ｐ$_d$）はカタカナの単語の“＾”及び“＄”以外の文字数である。なお、（３）式におけるｅｕ$_0^{n(Pd)+1}$は、ｅｕ$_0$からｅｕ$_{n(Pd)+1}$までの文字列を示し、（４）式におけるｋｕ$_0^{n(Pd)+1}$は、ｋｕ$_0$からｋｕ$_{n(Pd)+1}$までの文字列を示している。
【００３１】
Ｌ｛Ｋ｝中の各Ｐ$_d$における（４）式のｋｕ$_0^{n(Pd)+1}$が変換候補のカタカナ単語となる。また、（３）式のｅｕ$_0^{n(Pd)+1}$が（４）式の変換候補を出力する際の

英語の単語中の変換単位を示している。

【００３２】

変換候補生成手段１４は、他言語の単語等が格納されている変換候補生成規則１８を入力し、変換候補のラティス１９を出力する。出力された変換候補のラティス１９は、確率計算手段１５に入力される。

【００３３】

次に、確率計算手段１５は、入力された変換候補のラティス１９と、確率モデル１７とに基づいて、変換候補の生起確率を計算して選択手段１６に出力する。選択手段１６は、生起確率に基づいて変換候補を選択して出力する。また、確率計算手段１５の処理と選択手段１６の処理は交互に繰り返しながら少しずつ処理を行い、最適な変換候補として生起確率が最大となる変換候補を選択する。なお、選択手段１６は、確率計算手段１５による文脈情報を用いた変換候補の評価に基づいて変換候補の選択を行う。

【００３４】

ここで、変換候補の評価手法について説明する。まず、英語の単語を入力して対応するカタカナの単語Ｅ＾を推定するためには、以下に示す（６）式を満たすＫを求めればよい。

【００３５】

【数３】

$$E\hat{} = \arg\max_{K} P(K \mid E) \quad \cdots(6)$$

ここで、Ｐ（Ｋ｜Ｅ）は、Ｅが与えられた場合の、Ｋの条件付き確率分布を表す。しかしながら、（６）式を直接求めることは未知の単語に対して難しい。そこで、（３）式、（４）式により（６）式中の単語を部分文字列に分解する。分解した式を（７）式に示す。

【００３６】

【数４】

$$E\hat{} = \arg\max_{K} \sum_{eu_0^{n(P_d)+1}} \sum_{ku_0^{n(P_d)+1}} P(K \mid ku_0^{n(P_d)+1}) P(ku_0^{n(P_d)+1} \mid eu_0^{n(P_d)+1}) P(eu_0^{n(P_d)+1} \mid E) \quad \cdots(7)$$

（７）式では、ラティス上の同じＫを示す全ての変換候補の確率を合計することで、結果が得られることを示している。

【００３７】

また、（７）式のＰ（$eu_0^{n(Pd)+1}$｜Ｅ）は、英語の単語から生成される部分文字列の確率分布であり、変換単位推定モデルと呼ぶ。また、Ｐ（$ku_0^{n(Pd)+1}$｜$eu_0^{n(Pd)+1}$）は、英語の部分文字列から生成されたカタカナの部分文字列の確率分布であり翻訳モデルと呼ぶ。更に、Ｐ（Ｋ｜$ku_0^{n(Pd)+1}$）は、カタカナの部分文字列からカタカナの単語が生成される確率分布である。

【００３８】

ここで、上述の変換単位推定モデル、翻訳モデル、及び確率分布である式、Ｐ（Ｋ｜$ku_0^{n(Pd)+1}$）Ｐ（$ku_0^{n(Pd)+1}$｜$eu_0^{n(Pd)+1}$）Ｐ（$eu_0^{n(Pd)+1}$｜Ｅ）に実際の値を入力した例を図に示す。図３は、変換単位推定モデル、翻訳モデル、及び確率分布に実際の値を適用した一例の図である。なお、図３では、“ア

クチニウム（ａｃｔｉｎｉｕｍ）”を変換単位推定モデル、翻訳モデル、及び確率分布を示す式に適用し、「変換単位推定モデル×翻訳モデル×確率分布」を示している。

【００３９】

ここで、（７）式のＰ（ｋｕ$_0^{n(Pd)+1}$｜ｅｕ$_0^{n(Pd)+1}$）を、単語単位の処理から部分文字列単位の処理に分解する。これにより、下記に示す（８）式のようになる。

【００４０】

【数５】

$$P(ku_0^{n(p_d)+1} \mid eu_0^{n(p_d)+1}) = \prod_{i=0}^{n(p_d)+1} P(ku_i \mid ku_0^{i-1}, eu_0^{n(p_d)+1}) \quad \cdots(8)$$

更に、Ｐ（ｋｕ$_i$｜ｋｕ$_0^{i-1}$，ｅｕ$_0^{n(Pd)+1}$）の条件の英語の文字列をｅｕ$_i$と、ｅｕ$_i$の前ａ文字、ｅｕ$_i$の後ｂ文字だけに近似し、カタカナの文字列をｋｕ$_i$の前ｃ文字だけに近似する。

【００４１】

【数６】

$$\prod_{i=0}^{n(p_d)+1} P(ku_i \mid ku_0^{i-1}, eu_0^{n(p_d)+1}) \approx \prod_{i=0}^{n(p_d)+1} P(ku_i \mid k_{start\_ku(i)-c}^{start\_ku(i)-1}, e_{start\_eu(i)-a}^{start\_eu(i)-1}, eu_i, e_{start\_eu(i)+1}^{start\_eu(i)+b}) \quad \cdots(9)$$

ここで、ｓｔａｒｔ＿ｅｕ（ｉ）は、ｉ番目の部分文字列ｅｕ$_i$の初めの文字の位置を示し、ｓｔａｒｔ＿ｋｕ（ｉ）はｉ番目の部分文字列ｋｕ$_i$の初めの文字の位置を示している。また、上述のａ，ｂ，ｃは定数を示している。

【００４２】

（７）式の変換単位推定モデルＰ（ｅｕ$_0^{n(Pd)+1}$｜Ｅ）の確率は、Ｅ＝ｅｕ$_0^{m+1}$の単語を部分文字列に分割する確率（分割確率）であるので、各文字の間が分割点にあるかどうかで全ての分割パターンを表現することができる。分割可能な部分はｍ＋１個あり、それらが分割点かそうでないかの２値を取ることにより、全ての部分文字列への分割を表現することができる。ここで、ｅ$_j$とｅ$_{j+1}$との間が部分文字列の分割になるかどうかをＺ$_j$で表現する。

【００４３】

【数７】

$$z_j = \begin{cases} \text{分割点の場合} \\ \text{分割点ではない場合} \end{cases} \quad \cdots(10)$$

Ｚ$_j$を用いて、Ｐ（ｅｕ$_0^{n(Pd)+1}$｜Ｅ）を単語単位の処理から文字単位の処理に分解する。

【００４４】
【数８】

$$P(eu_0^{n(p_d)+1} \mid E) = \prod_{j=0}^{m} P(z_j \mid z_0^{j-1}, e_0^{m+1}) \qquad \cdots(11)$$

更に、Ｚ$_j$の前ａ’文字と後ｂ’文字と、Ｚ$_j$の前のｃ’の分割情報（分割点か、又は分割点ではないかの情報）とを考慮するように近似する。
【００４５】
【数９】

$$\prod_{j=0}^{m} P(z_j \mid z_0^{j-1}, e_0^{m+1}) \approx \prod_{j=0}^{m} P(z_j \mid z_{j-c'}^{j-1}, e_{j-a'+1}^{j+b'}) \qquad \cdots(12)$$

（９）式、（１２）式を用いると、（７）式は次のようになる。
【００４６】
【数１０】

$$k^{\wedge} \approx \arg\max_{K} \sum_{eu_0^{n(p_d)+1}} \sum_{ku_0^{n(p_d)+1}} P(K \mid ku_0^{n(p_d)+1}) \prod_{i=1}^{n(p_d)+1} P(ku_i \mid k_{start\_ku(i)-c}^{start\_ku(i)-1}, e_{start\_eu(i)-a}^{start\_eu(i)-1}, eu_i, e_{start\_eu(i)+1}^{start\_eu(i)+b}) \prod_{j=0}^{m} P(z_j \mid z_{j-c'}^{j-1}, e_{j-a'+1}^{j+b'}) \qquad \cdots(13)$$

この（１３）式が、本発明における第１の評価式である。
【００４７】
また、（１３）式とは別の方法について説明する。（１３）式に示すように同じＫを出力するｋｕ$_0^{n(Pd)+1}$とｅｕ$_0^{n(Pd)+1}$とについての合計を取らずに、部分文字列の組み合わせを１つだけ選択するように近似する。ここでは、Ｋは、ｋｕ$_0^{n(Pd)+1}$としている。
【００４８】
【数１１】

$$k^{\wedge} \approx \arg\max_{K = ku_0^{n(p_d)+1}} \prod_{i=1}^{n(p_d)+1} P(ku_i \mid k_{start\_ku(i)-c}^{start\_ku(i)-1}, e_{start\_eu(i)-a}^{start\_eu(i)-1}, eu_i, e_{start\_eu(i)+1}^{start\_eu(i)+b}) \prod_{j=0}^{m} P(z_j \mid z_{j-c'}^{j-1}, e_{j-a'+1}^{j+b'}) \qquad \cdots(14)$$

この（１４）式が本発明における第２の評価式である。
【００４９】
なお、（１３）式、（１４）式に示す評価式では、ｅｕ$_i$に対応するｋｕ$_i$の確率を求める際に、ｅｕ$_i$の前ａ文字及びｅｕ$_i$の後ｂ文字の英語の文脈と、ｋｕ$_i$の前ｃ文字の日本語の文脈情報を考慮している。これによって、ｅｕ$_i$の発音を示すｋｕ$_i$の推定精度

を向上させることができる。また、英単語を部分文字列に分割する際に、分割候補の部分の前ａ’文字とｂ’文字、前ｃ’の分割情報という文脈情報を考慮している。

【００５０】

このように、文脈情報を用いて、元の単語を部分文字列へ分割する確率と、元の単語の部分文字列をカタカナの部分文字列へ変換する確率とからカタカナの単語の生起確率を計算し、ビタビアルゴリズム（Ｖｉｔｅｒｂｉ　ａｌｇｏｒｉｔｈｍ）を利用して効率的に確率が最大となるカタカナを選択することで、他言語の単語から高精度にカタカナの単語へ変換することができる。つまり、ラティスＬ｛Ｋ｝中の経路から、本発明における第１の評価式又は第２の評価式を満たすカタカナの文字列を選択し、カタカナの単語として出力する。

【００５１】

＜確率モデルの生成における素性関数の定義内容＞

次に、モデル生成手段１２にて行う確率モデルの生成における素性関数の定義内容について説明する。

【００５２】

本発明における評価式（（１３）式，（１４）式）で文脈を考慮する場合には、最大エントロピー法に基づいて構築した確率モデルを利用する。この確率モデルを利用するとモデルが対応できるデータが過疎になることを避けながら文脈情報を全て考慮して確率を求めることができる。なお、本発明におけるモデルの生成においてはこの限りではなく、他の統計的手法を用いてもよい。また、以下の説明では、最大エントロピー法による学習を例として扱う。

【００５３】

まず、（１３）式と（１４）式中の翻訳モデルである（１５）式で利用する素性関数について説明する。

【００５４】

【数１２】

$$P(ku_i \mid k^{start\_ku(i)-1}_{start\_ku(i)-c}, e^{start\_eu(i)-1}_{start\_eu(i)-a}, eu_i, e^{start\_eu(i)+b}_{start\_eu(i)+1}) \quad \cdots(15)$$

ここでは、データが過疎になりにくいように文字情報だけでなく、子音、母音、半母音の区別の情報も利用する。そこで、ｅ$_j$の子音、母音、半母音の区別の情報をＧ（ｅｊ）と表す。

【００５５】

【数１３】

$$G(e_j)=\begin{cases} vowel & (e_j=\{a,i,u,e,o\}) \\ semi-vouel & (e_j=\{h,y\}) \\ consonant & (else) \end{cases} \quad \cdots(16)$$

このｅｕ$_i$，ｅ$_j$，Ｇ（ｅ$_j$），ｋｕ$_i$を夫々１つの属性として、それらの属性の組み合わせにより、素性関数を定義する条件を作成する。

【００５６】

最大エントロピー法に基づいてモデルを構築する際に最も重要なことは、素性関数をど

のように定義するかという点にある。そこで、本発明では変換対象の部分文字列に距離が近いことと、連続していることが重要であると考え、以下の属性の組み合わせにより素性関数を定義する。

【００５７】

「ku_iとeu_i」、「ku_iと、eu_iと、eu_iの前あるいは後、又は前後のいくつかのe」、「ku_iと、eu_iと、eu_iの前あるいは後、又は前後のいくつかのG（e）」、「ku_iと$k_{start_ku(i)-1}$」

また、分割モデルである（１７）式では、次の組み合わせにより、素性関数を定義する。

【００５８】

【数１４】

$$P(z_j \mid z_{j-c'}^{j-1}, e_{j-a'+1}^{j+b'}) \quad \cdots(17)$$

「Z_jとe_jとe_{j+1}」、「Z_jとe_jとe_{j+1}とe_jの前のいくつかのZとe」、「Z_jとe_jとe_{j+1}とe_jの後のいくつかのZ」、「Z_jとe_jとe_{j+1}とe_jの前のいくつかのZとeと後のいくつかのe」

これにより、素性関数を定義することができるため、この素性関数を用いて最大エントロピー法による確率モデルを生成することができる。

【００５９】

なお、最大エントロピー法によるモデルの学習では、例えば、Ｂｅｒｇｅｒ（１９９６）の確率モデルの構築手法等を用いて（１３）式と（１４）式で用いる（１５）式、又は（１７）式の確率モデルを、学習データ１１を用いて構築することができる（Ａｄａｍ Ｌ．Ｂｅｒｇｅｒ，Ｓｔｅｐｈｅｎ Ａ．Ｄｅｌｌａ Ｐｉｅｔｒａ，ａｎｄ Ｖｉｎｃｅｎｔ Ｊ．Ｄｅｌｌａ Ｐｉｅｔｒａ．１９９６．Ａ Ｍａｘｉｍｕｍ Ｅｎｔｒｏｐｙ Ａｐｐｒｏａｃｈ ｔｏ Ｎａｔｕｒａｌ Ｌａｎｇｕａｇｅ Ｐｒｏｃｅｓｓｉｎｇ．Ａｓｓｏｃｉａｔｉｏｎ ｆｏｒ Ｃｏｍｐｕｔａｔｉｏｎａｌ Ｌｉｎｇｕｉｓｔｉｃｓ，Ｖｏｌ．２２，Ｎｏ．１，ｐｐ．３９－７１．）。

【００６０】

これにより、本発明における素性関数を用いて最大エントロピー法による確率モデルを生成することができる。

【００６１】

ここで、上述したように翻字装置における専用の装置構成により本発明における翻字を行うこともできるが、上述した内容をコンピュータに実行させることができる実行プログラム（翻字プログラム）を生成し、例えば、汎用のパーソナルコンピュータ、ワークステーション等に翻字プログラムをインストールすることにより、本発明における翻字が実現可能となる。

【００６２】

ここで、本発明における実行可能なコンピュータのハードウェア構成例について図を用いて説明する。図４は、本発明における翻字処理が実現可能なハードウェア構成の一例を示す図である。

【００６３】

図４におけるコンピュータ本体には、入力装置３１と、出力装置３２と、ドライブ装置３３と、補助記憶装置３４と、メモリ装置３５と、各種制御を行うＣＰＵ（Ｃｅｎｔｒａｌ Ｐｒｏｃｅｓｓｉｎｇ Ｕｎｉｔ）３６と、ネットワーク接続装置３７とを有するよう構成されており、これらはシステムバスＢで相互に接続されている。

【００６４】

入力装置３１は、使用者が操作するキーボード及びマウス等のポインティングデバイスを有しており、使用者からのプログラムの実行等、各種操作信号を入力する。出力装置３２は、本発明における翻字処理を行うためのコンピュータ本体を操作するのに必要な各種ウィンドウやデータ等を表示するモニタを有し、ＣＰＵ３６が有する制御プログラムに基づいて実行結果等を表示することができる。

【００６５】

ここで、本発明において、コンピュータ本体にインストールされる実行プログラムは、例えば、ＣＤ－ＲＯＭ等の記録媒体３８等により提供される。プログラムを記録した記録媒体３８は、ドライブ装置３３にセット可能であり、記録媒体３８に含まれる実行プログラムが、記録媒体３８からドライブ装置３３を介して補助記憶装置３４にインストールされる。

【００６６】

補助記憶装置３４は、ハードディスク等のストレージ手段であり、本発明における実行プログラムや、コンピュータに設けられた制御プログラムの他に、ドライブ装置３３から読み取ることができる学習データや、学習フェーズにおいて作成された確率モデル１７や変換候補生成規則１８を蓄積し必要に応じて入出力を行うことができる。

【００６７】

ＣＰＵ３６は、ＯＳ（Ｏｐｅｒａｔｉｎｇ　Ｓｙｓｔｅｍ）等の制御プログラム、メモリ装置３５により読み出され格納されている実行プログラムに基づいて、各種演算や各ハードウェア構成部とのデータの入出力等、コンピュータ全体の処理を制御して、上述した翻字における各処理を実現することができる。プログラムの実行中に必要な各種情報は、補助記憶装置３４から取得することができ、また格納することもできる。

【００６８】

ネットワーク接続装置３７は、通信ネットワーク等と接続することにより、実行プログラムを通信ネットワークに接続されている他の端末等から取得したり、翻字手順を規定したプログラムを実行することで得られた実行結果又は本発明における実行プログラム自体を他の端末等に提供することができる。

【００６９】

上述したようなハードウェア構成により、特別な装置構成を必要とせず、低コストで高精度な翻字処理を実現できる。

【００７０】

次に、上述したようなハードウェア構成により実行される翻字プログラムにおける処理手順について、フローチャートを用いて説明する。なお、フローチャートは学習フェーズと実行フェーズとに分けて示しており、翻字プログラムは、ＣＰＵ３６により図４に示す各構成部を用いた後述の各処理手順を実行する。つまり、ユーザから入力装置３１を用いて翻字プログラムの実行指示が入力されると、補助記憶装置３４に格納されている翻字プログラムをメモリ装置３５に格納する。ＣＰＵ３６は、メモリ装置３５に格納された翻字プログラムにしたがって本発明における翻字処理に係る機能を実行する。

【００７１】

図５は、本発明の学習フェーズにおけるモデル生成手順を示す一例のフローチャートである。まず、学習データを入力する（Ｓ０１）。この学習データは、対訳の単語内において対応付けがされているデータとなる。次に、確率モデルを生成する（Ｓ０２）。ここでは、上述した素性関数を用い、（１３）式、（１４）式中の確率モデルを統計的に求める。その後、Ｓ０２により生成した確率モデルを出力する（Ｓ０３）。

【００７２】

次に、図６に、本発明の学習フェーズにおける変換候補生成規則作成手順の一例のフローチャートを示す。図６に示す変換候補生成規則作成処理では、まず、学習データを入力する（Ｓ１１）。この学習データは、上述した確率モデル生成手順にて使用される学習データと同様であり、対訳の単語内において対応付けがされているデータである。次に、入力した学習データに基づいて変換候補生成規則を作成する（Ｓ１２）。ここでは、部分対

応付けされたカタカナと英語の単語対を用いてカタカナへの翻字処理のための変換候補生成規則を作成する。その後、Ｓ１２にて生成された変換候補生成規則を出力する（Ｓ１３）。

【００７３】

次に、実行フェーズについて図を用いて説明する。図７は、本発明の実行フェーズにおける翻字手順を示す一例のフローチャートである。

【００７４】

図７において、まず、翻字を行うために他言語データが入力されると（Ｓ２１）、変換候補生成規則手順にて生成された変換候補生成規則を入力する（Ｓ２２）。次に、他言語データと変換候補生成規則とから変換候補を生成する（Ｓ２３）。具体的には、英語からカタカナへ変換する場合は、変換元となる英単語から変換候補生成規則を用いて変換先のカタカナの部分文字列からなるカタカナの変換候補のラティスＬ｛Ｋ｝を生成する。

【００７５】

次に、上述したモデル生成手順にて生成した確率モデルを入力する（Ｓ２４）。確率モデルを入力後、Ｓ２３にて生成した変換候補のラティスＬ｛Ｋ｝を対象に、文脈情報を用いて元の単語を部分文字列へ分割する確率、及び元の単語の部分文字列をカタカナの部分文字列へ変換する確率から変換候補となるカタカナの単語の生起確率を計算する（Ｓ２５）。

【００７６】

次に、最適な変換候補として、Ｓ２５にて計算された生起確率が最大となる変換候補を選択して出力する（Ｓ２６）。具体的には、（１３）式、（１４）式に示した評価式を満たす最適なカタカナの文字列を選択し、その文字列をカタカナの単語として出力する。

【００７７】

ここで、本発明における第２の評価式である（１４）式の上位解は、ダイナミックプログラミング（動的計画法）に基づく、最適な状態遷移が生じた場合の出力確率を求めるアルゴリズムであるビタビアルゴリズムによって、効率的に求めることができる。

【００７８】

また、第１の評価式の（１３）式を満たす解は、（１４）式の上位解となる経路のみを取り扱うことにより、高精度な近似解を効率よく求めることができる。

【００７９】

これにより、他言語の単語から高精度にカタカナの単語へ翻字することができる。また、実行プログラムを用いることで、特別な装置構成を必要とせず、汎用のコンピュータで本発明における翻字処理を実行できるため、低コストで高精度なカタカナへの翻字を実現することができる。

【００８０】

上述したように本発明によれば、他言語の単語からカタカナの単語への高精度な翻字を実現することができる。これにより、例えば、辞書に登録がない場合でも翻字処理によって外国語の単語からカタカナの単語を生成することができるため、外国から日本語へ機械翻訳する際の翻訳率を向上することができる。

【００８１】

以上本発明の好ましい実施例について詳述したが、本発明は係る特定の実施形態に限定されるものではなく、特許請求の範囲に記載された本発明の要旨の範囲内において、種々の変形、変更が可能である。

【図面の簡単な説明】

【００８２】

【図１】本発明における翻字装置の機能構成の一例を示す図である。

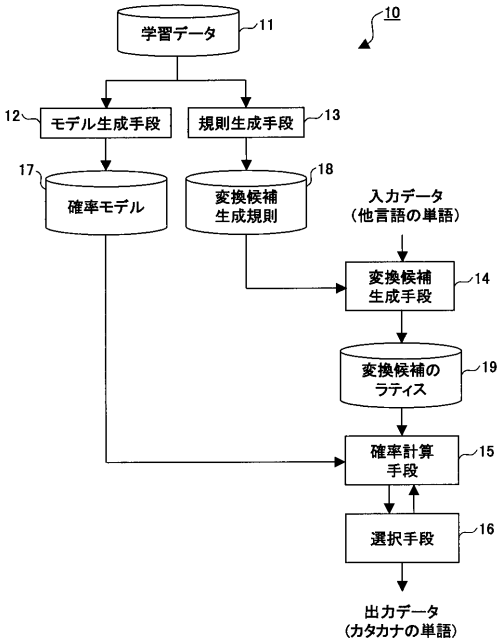
【図２】「ａｃｔｉｎｉｕｍ」の変換候補のラティスＬ｛Ｋ｝の例を示す。

【図３】変換単位推定モデル、翻訳モデル、及び確率分布に実際の値を適用した一例の図である。

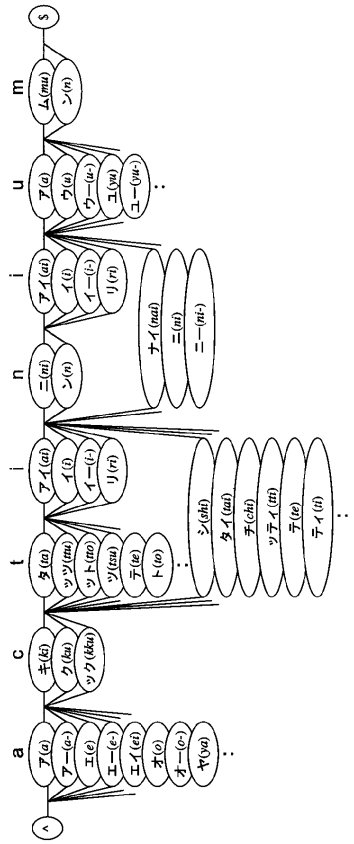
【図４】本発明における翻字処理が実現可能なハードウェア構成の一例を示す図である。

【図５】モデル生成手順を示す一例のフローチャートである。
【図６】変換候補生成規則作成手順を示す一例のフローチャートである。
【図７】翻字手順を示す一例のフローチャートである。
【符号の説明】
【００８３】
１０　翻字装置
１１　学習データ
１２　モデル生成手段
１３　規則生成手段
１４　変換候補生成手段
１５　確率計算手段
１６　選択手段
１７　確率モデル
１８　変換候補生成規則
１９　変換候補のラティス
３１　入力装置
３２　出力装置
３３　ドライブ装置
３４　補助記憶装置
３５　メモリ装置
３６　ＣＰＵ
３７　ネットワーク接続装置
３８　記録媒体

【図１】



【図２】



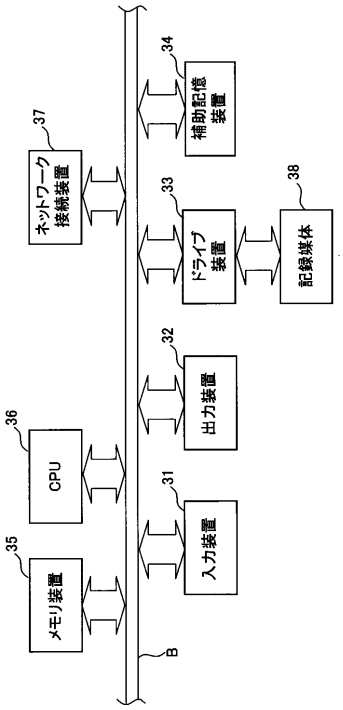
【図３】

P(^アクチニウム$|^/ア/ク/チ/ニ/ウ/ム/$)×
(a ku chi ni u mu) (a/ ku/ chi/ ni/ u/ mu)

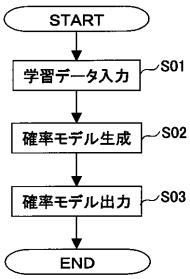
P(^/ア/ク/チ/ニ/ウ/ム/$|^/a/ku/chi/ni/u/mu/$)×
(a/ ku/ chi/ ni/ u/ mu)

P(^/a/c/ti/ni/u/m/$|^actinium$)

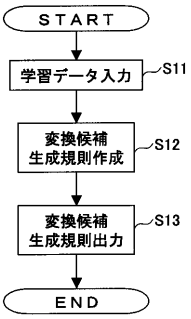
【図４】



【図５】



【図６】



【図７】