(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3921543号
(P3921543)
(45)発行日　平成19年5月30日(2007.5.30)　(24)登録日　平成19年3月2日(2007.3.2)

(51)Int.Cl.　F I
G06F 17/28 (2006.01)　G06F 17/28　Q

請求項の数 3　(全 20 頁)

(21)出願番号　特願2004-23914(P2004-23914)
(22)出願日　平成16年1月30日(2004.1.30)
(65)公開番号　特開2005-216127(P2005-216127A)
(43)公開日　平成17年8月11日(2005.8.11)
審査請求日　平成16年1月30日(2004.1.30)

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100130111
弁理士　新保　斉
(74)代理人　100090893
弁理士　渡邊　敏
(72)発明者　内元　清貴
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内
(72)発明者　井佐原　均
東京都小金井市貫井北町４－２－１　独立行政法人通信総合研究所内

審査官　和田　財太

最終頁に続く

(54)【発明の名称】機械翻訳装置

(57)【特許請求の範囲】

【請求項１】

所定の翻訳元言語で記述される翻訳元テキストを翻訳先言語で記述される翻訳先テキストに機械翻訳する機械翻訳装置であって、

翻訳元テキストを受理する受理手段、

該翻訳元テキストを形態素解析してその結果得られた各形態素を順に着目タームとし、翻訳元言語で記述された複数の文書を含む翻訳元言語文書データベースを用いて、該文書データベースの中から該着目タームが含まれる文書と、該文書データベースに含まれる全文書とのそれぞれにおける着目タームの分布間の距離を算出し、該距離が所定の閾値以上の着目タームを特徴的な意味を有する特徴語として抽出する特徴語抽出手段、

該特徴語を、翻訳先言語で表現される訳語に翻訳する訳語選択手段、

該訳語を含む文又は語句の少なくともいずれかを、翻訳先言語で記述された複数の文章を含む翻訳先言語データベースから抽出し、当該訳語と、文又は語句の少なくともいずれかとの関係を自動獲得した生成規則に基づいて文字単位候補を生成する文字単位候補生成手段、

生成された全ての文字単位候補間で依存関係が成立しうる文字単位候補対を全ての文字単位候補について抽出することを繰り返し、異なる依存関係で構成された翻訳先テキスト候補を生成する翻訳先テキスト候補生成手段、

各翻訳先テキスト候補の評価値を算出する評価手段、

該評価値に関連して、少なくとも翻訳先テキスト候補のうち１つを出力する出力手段

を備えたことを特徴とする機械翻訳装置。
【請求項２】
所定の翻訳元言語で記述される翻訳元テキストを翻訳先言語で記述される翻訳先テキストに機械翻訳する機械翻訳装置であって、
翻訳元テキストを受理する受理手段、
該翻訳元テキストを形態素解析してその結果得られた各形態素から、単語列の主辞となる形態素のうち、品詞が動詞、形容詞、名詞、指示詞、副詞、接続詞、連体詞、感動詞のいずれかである語を特徴語として抽出する特徴語抽出手段、
該特徴語を、翻訳先言語で表現される訳語に翻訳する訳語選択手段、
該訳語を含む文又は語句の少なくともいずれかを、翻訳先言語で記述された複数の文章を含む翻訳先言語データベースから抽出し、当該訳語と、該文又は語句の少なくともいずれかとの関係を自動獲得した生成規則に基づいて文字単位候補を生成する文字単位候補生成手段、
生成された全ての文字単位候補間で依存関係が成立しうる文字単位候補対を全ての文字単位候補について抽出することを繰り返し、異なる依存関係で構成された翻訳先テキスト候補を生成する翻訳先テキスト候補生成手段、
各翻訳先テキスト候補の評価値を算出する評価手段、
該評価値に関連して、少なくとも翻訳先テキスト候補のうち１つを出力する出力手段
を備えたことを特徴とする機械翻訳装置。
【請求項３】
前記機械翻訳装置において、
前記特徴語抽出手段で抽出された特徴語のうち、該特徴語間の依存関係情報を抽出する依存関係抽出手段を備え、
該依存関係情報を有する特徴語から生成された文字単位候補については、前記翻訳先テキスト候補生成手段において、該依存関係情報を用いて翻訳先テキスト候補を生成する
ことを特徴とする請求項１又は２に記載の機械翻訳装置。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は機械翻訳装置に関する。特に、翻訳元言語のテキストを入力し、翻訳先言語のテキストを出力するまでの処理手法に係る技術である。
【背景技術】
【０００２】
従来の機械翻訳方法においては、例えば特許文献１に開示される機械翻訳の手法が知られている。該特許文献でも開示されているように従来の機械翻訳方法は、例えば日本語から英語への翻訳を行う場合に、日本語の解析を行って、文の意味構造を導出する。
この意味構造は木構造や意味ネットワークで表現できる。そして、生成された意味ネットワークを解析し、適切な訳語を選択して目的とする訳文を生成する処理を行っているものがある。
【０００３】
ここで、意味ネットワークは意味記号を付加したノードと、ノード間の関係情報を付加したアークとによって記述されており、各ノードには翻訳文を生成する際に必要な文生成規則が付加される。生成時には文生成規則に基づき、英語の単語とその文法属性とを検索し、次にアークの情報と整合する単語、品詞、文型をそれぞれ決定する処理を行う。
【０００４】
【特許文献１】特公平８-３３８９５号公報
【０００５】
また、近年、大量のコーパスが利用可能となり、自然な表層文を生成する目的にもコーパスが利用されるようになってきている。その典型例の１つが機械翻訳に用いられる言語モデルであり、例えば非特許文献１で開示されている。

ここで示される統計的機械翻訳では、原言語で書かれたテキストをＳ、目的言語で書かれたテキストをＴとして、Ｓが与えられたときにＴを生成する確率Ｐが最大になるようなテキストを最適な翻訳として出力する。
この時用いられるモデルとしては、単語や句を原言語から目的言語に置き換えるための翻訳モデルと、置き換えた単語や句を目的言語側で尤もらしい順序に並べ替えるための言語モデルがある。
【０００６】
【非特許文献１】Brown,P.F., Cocke,J., Piera,S.A.D.,Pietra,V.J.D., Jelinek,F., Lalfferty,J.D., Mercer,R.L., and Roossin,P.S.(1990). "A Statistical Approach to MachineTranslation." Computationallingustics,16(2),79-85
【０００７】
このような従来の統計的機械翻訳の手法では、与えられた語の集合を並べ換えると自然な文を生成できるという仮定がある。つまり、自然な文を生成するための語の集合は翻訳モデルにより過不足なく生成されることが前提となっている。
しかし、この前提のためには大規模な対訳コーパスが必要であり、日本語などの比較的コーパスが整備された言語が原言語であっても、対象言語との対訳コーパスの状況、対象言語におけるコーパスの状況によっては、上記従来の手法では十分な翻訳精度の向上が望めない問題があった。
【０００８】
また、特に話し言葉や、文法的な誤り、記述上の誤りなど原言語の表現が不適切な場合に、これら従来の方法では正しい意味構造の解析や翻訳モデルの作用が不可能であり、全く誤った翻訳結果を出力してしまうことがある。
【発明の開示】
【発明が解決しようとする課題】
【０００９】
本発明は、上記従来技術の有する問題点に鑑みて創出されたものであり、その目的は、翻訳元言語又は翻訳先言語に十分なコーパス等の環境が整っていない場合や、入力するテキストが不正確な場合にも、好適な翻訳先言語のテキストを出力可能な機械翻訳装置を提供することである。
【課題を解決するための手段】
【００１０】
本発明は、上記の課題を解決するために、次のような機械翻訳装置を提供する。
すなわち、所定の翻訳元言語で記述される翻訳元テキストを翻訳先言語で記述される翻訳先テキストに機械翻訳する機械翻訳装置である。
そして、請求項１に記載の機械翻訳装置は、（１）翻訳元テキストを受理する受理手段、（２）該翻訳元テキストを形態素解析してその結果得られた各形態素を順に着目タームとし、翻訳元言語で記述された複数の文書を含む翻訳元言語文書データベースを用いて、該文書データベースの中から該着目タームが含まれる文書と、該文書データベースに含まれる全文書とのそれぞれにおける着目タームの分布間の距離を算出し、該距離が所定の閾値以上の着目タームを特徴的な意味を有する特徴語として抽出する特徴語抽出手段、（３）該特徴語を、翻訳先言語で表現される訳語に翻訳する訳語選択手段、（４）該訳語を含む文又は語句の少なくともいずれかを、翻訳先言語で記述された複数の文章を含む翻訳先言語データベースから抽出し、当該訳語と、該文又は語句の少なくともいずれかとの関係を自動獲得した生成規則に基づいて文字単位候補を生成する文字単位候補生成手段、（５）生成された全ての文字単位候補間で依存関係が成立しうる文字単位候補対を全ての文字単位候補について抽出することを繰り返し、異なる依存関係で構成された翻訳先テキスト候補を生成する翻訳先テキスト候補生成手段、（６）各翻訳先テキスト候補の評価値を算出する評価手段、（７）該評価値に関連して、少なくとも翻訳先テキスト候補のうち１つを出力する出力手段を備えたことを特徴とする。
【００１１】

請求項２に記載の機械翻訳装置は、上記請求項１の機械翻訳装置において、特徴語抽出手段が翻訳元テキストを形態素解析してその結果得られた各形態素から、単語列の主辞となる形態素のうち、品詞が動詞、形容詞、名詞、指示詞、副詞、接続詞、連体詞、感動詞のいずれかである語を特徴語として抽出する構成である。

【００１２】

請求項３に係る機械翻訳装置は、前記特徴語抽出手段で抽出された特徴語のうち、該特徴語間の依存関係情報を抽出する依存関係抽出手段を備え、該依存関係情報を有する特徴語から生成された文字単位候補については、前記翻訳先テキスト候補生成手段において、該依存関係情報を用いて翻訳先テキスト候補を生成することができる。

【発明の効果】

【００１３】

以上の発明により次の効果を奏する。

すなわち請求項１ないし３に記載の機械翻訳装置によれば、新しい機械翻訳手法を導入することで、翻訳元言語と翻訳先言語に十分なコーパス（特に対訳コーパス）が整っていない場合でも、高精度な機械翻訳装置を提供することができる。

また、入力するテキストから特徴語を抽出するため、文法的に不正確なテキストを入力しても、自然で理解のしやすい翻訳を実現することができる。

【００１４】

特に、請求項３に記載の機械翻訳装置では、翻訳元言語テキストにおける依存関係を、翻訳先言語の生成ステップで用いることができるので、高精度な翻訳に寄与することができる。

【００１５】

このような機械翻訳装置は、特に口語を入力したものや、記事のタイトルなど文法的に正確でない翻訳元言語の文章を翻訳する際に適している。得られる結果は大意が分かりやすい翻訳先テキストである。

また、本実施例で示した日本語・英語間のようなコーパス環境の整った言語間はもとより、地域性の強い言語や、発展途上国の固有言語など、コーパスが未整備な言語に翻訳する際にも、実際的な機械翻訳結果を生成することができる。

【発明を実施するための最良の形態】

【００１６】

以下、本発明の最良と考えられる実施形態を、図面に示す実施例を基に説明する。なお、実施形態は下記に限定されるものではない。

図１には本発明による第１実施例の機械翻訳方法のフローチャートを示す。図のように、翻訳元言語である日本語のテキスト（１）を入力受理し、該テキスト（１）中の特徴的単語を抽出（２）する。抽出された特徴的単語は特徴的単語データ（６）として記録する。なお、本発明はいかなる言語をも対象とするので、本発明は単語として分類できるものに限らず、単語、語句など広く含む特徴語を抽出し、後段の処理を行ってもよい。

【００１７】

次に、翻訳元言語と翻訳先言語の辞書データ（７）を用いて特徴的単語の最適な訳語を選択（３）する。これにより、翻訳先言語の特徴的単語が生成され特徴的単語訳語データ（８）として記録する

さらに、特徴的単語訳語データ（８）を用い、翻訳先言語のテキストを生成（４）する。このように特徴的単語訳語から、自然な英語テキスト（５）を生成することで機械翻訳を行うプロセスは、従来創出されていない、本発明の最大の特徴である。

【００１８】

テキスト生成（４）は、特徴的単語訳語から所定の生成規則に基づいて文字単位候補を生成し、その各文字単位間の依存関係の組み合わせをかえて翻訳先テキスト候補を生成する。さらに、翻訳先テキスト候補を評価して最も適当と判断される英文テキスト（５）を出力する。

【００１９】

また、本発明では図２に示すような第２実施例の機械翻訳方法を提供することができる。翻訳元言語である日本語のテキスト（１０）を入力し、該テキスト（１０）中の特徴的な単語を抽出（１１）する。抽出された特徴的単語は特徴的単語データ（１６）として記録する。

【００２０】

特徴的単語抽出時に、各単語間の依存関係を抽出（１２）する。該単語間依存関係データ（１７）を記録する。

そして、翻訳元言語と翻訳先言語の辞書データ（１８）を用いて特徴的単語の最適な訳語を選択（１３）する。これにより、翻訳先言語の特徴的単語が生成され特徴的単語訳語及び前述の依存関係のデータ（１９）を記録する

さらに、該データ（１９）を用い、翻訳先言語のテキストを生成（１４）する。このように特徴的単語訳語から、自然な英語テキスト（１５）を生成する。

【００２１】

テキスト生成（14）は、特徴的単語訳語から所定の生成規則に基づいて文字単位候補を生成し、その各文字単位間の依存関係を仮定して単数又は複数の翻訳先テキスト候補を生成する。上記で特徴的単語の依存関係が記録されている場合には、その依存関係を元に翻訳先テキスト候補を生成することで、依存関係に誤りのない高精度な翻訳先テキストが生成できる。

さらに、翻訳先テキスト候補を評価して最も適当と判断される英文テキスト（５）を出力する。

【００２２】

本発明の第３実施例としては、上記第２実施例のテキスト生成（１４）における処理を一部変更し、記録した単語間依存関係（１７）に関わらず、特徴的単語訳語から所定の生成規則に基づいて文字単位候補を生成し、その各文字単位間の依存関係を仮定して単数又は複数の翻訳先テキスト候補を生成する。

その上で、翻訳先テキスト候補を評価する際に、上記単語間依存関係との同一性を調べ、その値が高い英語テキストを出力する。

【００２３】

以上に述べたような機械翻訳方法を実現する機械翻訳装置の構成を図３に示す。本装置（３０）は、例えばある新聞記事「官邸前などでドゥダエフ政権部隊と激しい市街戦を展開している。」という日本語テキスト（３１）を入力すると、受理手段である入力受理部（４０）で装置（３０）内への取り込み処理を行い、特徴的単語抽出部（５０）において「官邸」「ドゥダエフ」「部隊」「激しい」「戦」「展開」を日本語テキスト（３１）から抽出する。

【００２４】

さらに訳語選択部（６０）で各特徴的単語の最適な訳語、ここでは「palace」「Dudayev」、「troops」、「fierce」、「battle」、「engage」を選択する。

選択された特徴的単語訳語を、テキスト生成部（７０）において適切な単語間の補完を行いながらテキスト生成（３５）し、出力部（８０）から英語テキスト（３２）を出力する。

次に各部（４０）ないし（８０）の詳細を説述する。

【００２５】

入力受理部（４０）は図４に示すようにＣＰＵ（４１）とそれに接続されたスキャナ（４２）や、ＣＤドライブ、ハードディスクドライブ、ＭＯドライブ、フロッピー（登録商標）ディスクドライブなどの外部記憶装置（４３）等から構成される。また、ＣＰＵ（４１）の動作に伴い、必要に応じて公知のメモリを用いることもできる。

スキャナ（４２）を備える場合にはＣＰＵ（４１）において文字認識処理を行い、テキストデータに変換して外部記憶装置（４３）に記録する。外部記憶装置（４３）から直接日本語テキスト（３１）のデータを読み出す場合にも、ＣＰＵ（４１）において本装置（３０）で処理可能な形式にデータ変換を行うこともできる。

【００２６】
また、本発明はインターネットやイントラネットのネットワーク（４４）を介して他のコンピュータサーバー等からテキストデータを受信することも可能である。
入力受理部（４０）により日本語テキスト（３１）は図５に示される特徴的単語抽出部（５０）に送られる。
【００２７】
特徴的単語抽出部（５０）の構成を図５に示す。ここでもＣＰＵ及びメモリが協働して各処理を行う。本特徴的単語抽出部（５０）では、入力された日本語テキスト（３１）からそのテキストの内容を特徴的に表す特徴語を抽出する。
このような技術は、言語処理において文書を要約する技術や、文書検索などの要素技術として公知の多数の手法が知られており、それらを適宜用いることができるが、ここでは一例として非特許文献２に記載の方法を用いる。
【００２８】
【非特許文献２】情報処理学会自然言語処理研究会 １９９９-ＮＬ-３３，１９９９「タームのrepresentativeness」を測る」 久光徹、丹羽芳樹、辻井潤一
【００２９】
本方法によると、特徴語を選ぶために文書中の単語の話題性もしくは分野代表性（representativeness、本明細書ではこれを特徴性と呼ぶ。）を測ることが可能であり、かつ数値的な評価によるため、本発明の実施に好適である。以下に、簡単に説述する。
まず、本特徴的単語抽出部（５０）では、公知の形態素解析技術を用いて、日本語テキスト（３１）を形態素解析部（５１）において形態素解析する。解析された形態素はメモリ又は図示しない外部記憶装置などに形態素テーブルとして記録する。
【００３０】
そして、形態素テーブルから形態素を順次読み出し、その形態素（以下、これを着目タームと呼ぶ）毎に特徴性を測る。
まず文書抽出部（５２）において、着目タームWについて、Wを含む文書すべてを任意の文書データベース（５６）から抽出する。文書データベース（５６）は複数の日本語（翻訳元言語）の文書が含まれたものであり、外部記憶装置などに記憶されている。日本語単言語のコーパスや日英の対訳コーパスの日本語部分を用いてもよい。
【００３１】
次に、着目タームWが抽出された文書すべての集合における単語分布と、文書データベース（５６）に含まれる全文書の単語分布とを、単語分布算出部（５３）において算出し、各単語分布間の異なりの度合いを測る。
具体的には異なり度合算出部（５４）において次のような計算処理を行う。
【００３２】
すなわち、着目タームW、Wを含む文書すべての集合Ｄ（Ｗ）、全文書の集合D_0、Ｄ（Ｗ）における単語分布$P_{D(W)}$、D_0における単語分布P_0として、Wの特徴性Ｒｅｐ（Ｗ）を、２つの分布｛$P_{D(W)}$，P_0｝の距離Dist｛$P_{D(W)}$，P_0｝に基づいて定義する。
単語分布間の距離計測方法として、本実施例では対数尤度比を用いている。すなわち、全単語を｛W_1,…,W_n｝、単語w_iがＤ（Ｗ）、D_0に出現する頻度をそれぞれk_i、K_iとするとき、$P_{D(W)}$、P_0の距離Dist｛$P_{D(W)}$，P_0｝を、次のように定義する。
【００３３】
【数１】

$$Dist(P_{D(W)}, P_0) = \sum_{i=1}^{n} k_i \log \frac{k_i}{\#D(W)} \sum_{i=1}^{n} k_i \log \frac{K_i}{\#D_0}$$

ここで、＃Ｄ（Ｗ）は着目タームWについてＤ（Ｗ）の含む単語数、＃D_0は同様に全文

書の含む単語数である。
【００３４】
数１の定義によると、＃Ｄ（Ｗ）が離れた着目ターム同士の特徴性を有効に比較することが難しいため、数２のように正規化を行った特徴性Ｒｅｐ（Ｗ）を定義する。なおＢ（・）は＃Ｄ（Ｗ）が適切な数となる範囲内（例えば１０００≦＃Ｄ（Ｗ）≦２００００）で特徴性が精度よく求められるような指数関数を用いた近似関数である。
【００３５】
（数２）
$Rep(W) = Dist\{P_{D(W)}, P_0\} / B(\#D(W))$
【００３６】
ここで、「する」などのように著しく＃Ｄ（Ｗ）が大きい場合には、Ｄ（Ｗ）の抽出数を限定し、＃Ｄ（Ｗ）≦２００００を満たすようにすることで、上記近似関数を有効に用いることができると共に計算量を削減できる。
【００３７】
特徴的単語抽出部（５０）では以上の方法により特徴性を算出すると共に、所定の閾値に従って、特徴的単語決定部（５５）により入力した日本語テキスト（３１）の特徴的単語を抽出する。この結果は外部記憶装置に特徴的単語テーブル（３３）として保存される。
【００３８】
次に、図６に示す訳語選択部（６０）では、特徴的単語（３３）毎に最適な訳語、特徴的単語訳語（３４）を選択する処理を行う。単語の翻訳は、通常辞書データベースを参照することにより可能であるが、異なる言語間ではしばしば単語の多義性の問題が生じる。すなわち、ある単語を入力しただけでは、複数の訳語のいずれを選択するべきかが不明であり、最適な訳語の選択は本機械翻訳装置（３０）の翻訳精度に影響する極めて重要な技術である。
【００３９】
このような多義性解消については、公知の優れた技術が多数提案されており、本発明の実施には任意の技術を用いることができるが、ここでは本件出願人らによって非特許文献３及び非特許文献４に開示される次のような訳語選択モデルを用いる例を示す。
【００４０】
【非特許文献３】電子情報通信学会 言語理解とコミュニケーション研究会 NLC2001-41「翻訳メモリとコーパスを用いた学習に基づく訳語選択」内元清貴、関根聡、村田真樹、井佐原均 ２００１年
【非特許文献４】「Word Translation based on Machine LearningModels Using Translation Memory and Corpora」 KiyotakaUchimoto, Satoshi Sekine,Masaki Murata, and Hitoshi Isahara. SENSEVAL-2,pp.155-158,2001
【００４１】
訳語選択部（６０）は特徴的単語テーブル（３３）を入力として、まず文字列の類似性判定部（６１）において、文字列の類似性に基づく方法で訳語の決定を試みる。
すなわち、該テーブル（３３）に含まれる特徴的単語が、対訳辞書データベース（６３）において一義的に訳語が選択可能な場合には、該訳語を特徴的単語訳語テーブル（３４）に出力する。
【００４２】
また、日本語と英語の対訳コーパスデータベース（本実施例では対訳辞書との明確な区別を設けずに１個のデータベースとする）（６３）を参照し、特徴的単語テーブル（３３）との類似度を求める。
例えば、特徴的単語として「母」「遠慮」が含まれるとき、対訳コーパスデータベース（６３）に含まれる「母に遠慮」「母への遠慮」「献金を遠慮」などとの類似度を求める。
【００４３】
類似度を求める方法は、非特許文献５で開示されるＵＮＩＸ（登録商標）のｄｉｆｆコマンドを用いた方法によってもよく、以下の数３で求められる。

【００４４】
【非特許文献５】情報処理学会自然言語処理研究会 ２００１-ＮＬ-４４，２００１ 「ｄｉｆｆと言語処理」 村田真樹、井佐原均
【００４５】
（数３）
類似度＝（特徴的単語テーブルと対訳コーパスとのｄｉｆｆをとったときに一致した文字数）／（対訳コーパスの文字数）
【００４６】
ここで対訳コーパスについても、比較する前に機能語・動詞・形容詞の活用部分、サ変動詞をすべて削除する、あるいは該対訳コーパスによっても前述の特徴的単語の抽出を行い、それらの結果と類似度を比較するのが望ましい。
これにより、比較の対象として不適切な文字を多く含む対訳コーパスとの比較を避け、効果的な類似度の算出が可能となる。
【００４７】
類似度が所定の閾値（対訳コーパス及び特徴的単語テーブルの内容により適宜設定することができる）を超える対訳コーパスを特徴的単語の訳語を含む対訳コーパスとして選択する。
そして、該対訳コーパス中の訳語を特徴的単語訳語テーブル（３４）に記録する。対訳コーパス（６３）を用いるので、各単語の対訳関係はあらかじめ分かっており、特徴的単語と訳語の関係は機械的に決定できる。
【００４８】
ところで、対訳辞書・対訳コーパスデータベース（６３）に入力される全ての特徴的単語テーブルに類似する文例を備えることは難しく、類似度が閾値以上とならないものが多数残ってしまう場合が考えられる。
そこでさらに、機械学習モデルを用いた類似性判定部（６２）を備え、学習したデータ（６４）を用いながら最適な訳語の選択を行う。
【００４９】
機械学習のモデルとしては、ＳＶＭ（Support Vector Machine）を用いるが、ＭＥ（MaximumEntropy）、ＤＬ（Decision List）、ＳＢ(Simple Bayes)を用いてもよい。各分類クラスの確信度は基本的に文脈の集合をＢ、クラスの集合をＡとするとき、文脈ｂ（∈Ｂ）でクラスａ（∈Ａ）となる事象（a,b）の確率分布ｐ（a,b）として求められる。ＳＶＭではこのような確率分布は得られないが、便宜的に最適のクラスに対して確率値を１、その他のクラスに対して０とする。
【００５０】
文脈ｂの素性としては、例えば次のものを用いることができる。
（１）形態素情報
（２）文字ｎ-ｇｒａｍ
（３）最大一致となる用例に関する情報
（４）内容語とのその訳語候補の出現頻度
ＳＶＭによって、対訳コーパスを用いて機械学習を行い、学習データ（６４）として記録する。
【００５１】
特徴的単語について対訳辞書データベース（６３）から訳語候補を抽出する。そして、訳語候補の素性は英主辞単語、文集合、単語、出現頻度の組み合わせによって表される。機械学習による学習データ（６４）を用い、訳語候補から特徴的単語訳語を選択し、特徴的単語訳語テーブル（３４）に記録する。
このような訳語選択モデルには公知の手法を任意に用いることが可能であり、上記のような機械学習モデルによるものでなくともよい。
【００５２】
ここで、特徴的単語の訳語を選択する際に、固有表現などが抽出されて対訳辞書データベ

ースに訳語がない場合が想定される。このとき、訳語選択モデルに問い合わせ処理を行う処理部を設けてユーザーに問い合わせを行ってユーザが与えるようにしてもよい。
これを自動化する方法としては、非特許文献６に開示されるように、単に翻訳元言語の発音に従って、一定のルールにより翻訳先言語の文字に置き換える（音訳する）こともできるし、提案されているモデルを用いることもできる。ここで提案されているのは、人名や組織名などで、まず音訳を行ったり、適当な訳語で翻訳して訳語候補を作成し、その候補の中から所定のテキストデータベースに出現する頻度の高いものを訳語として選択するものである。テキストデータベースとしては、例えばインターネットのＷｅｂ情報などを利用することができる。

【００５３】

次に、以上により形成された特徴的単語訳語テーブル（３４）を、テキスト生成部（７０）に入力し、英語テキストを生成する。
いくつかの単語を入力し、その単語を含むテキストを生成する方法としては次のような手法がある。すなわち、本件出願人が特許文献２で開示するテキスト生成方法を、翻訳先言語である英語に適用して用いる。

【００５４】

【特許文献２】特開２００３-２７１５９２号公報

【００５５】

本テキスト生成部（７０）の具体的な構成例として図７に示す各部を備える。テキスト生成部（７０）は、例えばＣＰＵとメモリ、ハードディスクなどの外部記憶媒体を備えるパーソナルコンピュータなどにより構成することができ、主な処理をＣＰＵにおいて行い、処理の結果を随時メモリ、外部記憶媒体に記録する。

【００５６】

本実施例で、入力された特徴的単語訳語（３４）は２つの処理に用いられる。その１つは単語列生成規則獲得部（７１）であり、もう１つは単語列候補生成部（７２）である。
このとき、特徴的単語訳語（３４）は単語列の主辞となる内容語であると定義する。また、内容語は、その語の品詞が、動詞、形容詞、名詞、指示詞、副詞、接続詞、連体詞、感動詞、未定義語である形態素の見出し語であるとし、それ以外の形態素の見出し語を機能語とする。

【００５７】

単語列生成規則獲得部（７１）では、特徴的単語訳語（３４）が与えられたとき、それぞれを含む文を英語コーパス（７５）から検索し、形態素解析、構文解析(係り受け解析)をする。そして、そこから特徴的単語訳語（３４）を含む単語列を抽出して、特徴的単語訳語から単語列を生成する単語列生成規則（７６）を獲得し、記録する。例えば、「palace」→「atthe palace」、「palace」→「in the palace」、「battle」→「battles」、「engage」→「They areengaged in」、「engage」→「I engaged」などの単語列生成規則（７６）を獲得し、記録する。

【００５８】

ここで、生成規則の自動獲得には次の手法を用いる。特徴的単語訳語の集合をVとし、特徴的単語訳語k(∈V) から単語列を生成する規則の集合をRk とするとき、規則rk(∈Rk)は次の形式で表現されるものと定義する。

k → hk m*

hk は特徴的単語訳語を含む主辞形態素、m*は同じ単語列内でhkに連続する任意個の形態素とする。特徴的単語訳語が与えられると、この形式を満たす規則を翻訳先言語のコーパス（７５）から自動獲得する。

【００５９】

一方、単語列候補生成部（７２）では、単語列生成規則（７６）を参照しながら、入力された特徴的単語訳語（３４）から出力する英語テキスト（３２）を構成する単語列の候補を生成する。
例えば、「palace」では自然なテキストを構成する単語列とはなりにくいが、「at the p

alace」あるいは「inthe palace」のように「palace」という単語と極めて密接な関連性を有する語句を付加し、後段の処理によるテキスト生成に備える。

【００６０】

本実施例のように、単語列生成規則獲得部（７１）によりコーパス（７５）から入力する特徴的単語訳語（３４）の単語列規則を生成することで、最小限の計算量で効果的に単語列生成規則を得ることができ、処理速度の向上に寄与する。

【００６１】

もっとも、必ずしも特徴的単語訳語（３４）に関連する語句をコーパスから抽出する構成を取る必然性はなく、計算能力に応じて任意の語句を入力された特徴的単語訳語（３４）の前後に付加してもよい。あるいは、本装置（３０）内に備える対訳辞書データベース（６３）に含まれる慣用表現の情報から単語列を生成することもできる。上記「palace」→「at the palace」などは対訳辞書データベースに記載される表現であり、単語列の候補として生成することができる。

【００６２】

また、日本語など主格を多く省略する言語を入力した場合には、「engage」→「They are engagedin」などのように主語を補って単語列候補を生成することができる。このとき、日本語などの多くの言語では主格が明らかな時や、形式主語であるときに省略されることに着目し、入力に主格が何であるかの情報だけでなく、主格がないという情報を用いることで、「engage」→「Heis engaged in」を生成せず、「engage」→「They are engaged in」を生成するようにすることもできる。

【００６３】

次に、テキスト候補生成部（７３）においてテキスト候補を生成する。テキスト候補はグラフあるいは木の形で表現する。ここでは特徴的単語訳語（３４）のうち、「palace」「troops」「engage」の３語の関係を例として説述する。

すなわち、図１２のように、各単語列候補（３４ａないし３４ｆ）の間に係り受けの関係を仮定して、テキスト候補１（３５'）、テキスト候補２（３５''）のような単語列を単位とした依存構造木の形でテキスト候補を生成する。このとき、３語の場合に全ての係り受け関係はR！×２＝１２通りであるが、翻訳先言語の文法・特性に合わせて語順の固定などにより候補の数を削減することができる。

【００６４】

生成されたテキスト候補（３５'など）は、評価部（７４）でコーパスから学習した特徴的単語訳語生成モデル（７７）や言語モデル（７８）を用いて順序付けされる。

以下、特徴的単語訳語生成モデル（７７）と、言語モデル（７８）として形態素モデル及び係り受けモデルについて説述する。

【００６５】

特徴的単語訳語生成モデルでは、次の５種類の情報を素性として用いたモデル（ＫＭ１ないし５）を考える。以下で、特徴的単語訳語の集合Vは、ある回数以上コーパスに出現した主辞単語の集合とし、単語列は前記で表現されるものと仮定する。また、各特徴的単語訳語は独立であり、与えられたテキストが単語列w1・・・wmからなるとき、特徴的単語訳語ki は単語wj(1≦j≦m)に対応していると仮定する。図１３にモデルの説明図を示す。

【００６６】

[ＫＭ１]

前方の二単語を考慮(trigram)

ki は前方の二単語wj-1 とwj-2 のみに依存すると仮定する。

【数４】

$$P(K|M,D,T) = \prod_{i=1}^{m} P(k_i|w_{j-1},w_{j-2})$$

【００６７】
[ＫＭ２]
後方の二単語を考慮(後方trigram)
ki は後方の二単語wj+1 とwj+2 のみに依存すると仮定する。
【数５】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_{j-1},w_{j-2})$$

【００６８】
[ＫＭ３]
係り単語列を考慮(係り単語列)
ki を含む単語列に係る単語列がある場合、ki はそのうち最も文末側の単語列の末尾から二単語wl とwl-1 のみに依存すると仮定する(図１３参照)。
【数６】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_l,w_{l-1})$$

【００６９】
[ＫＭ４]
受け単語列を考慮(受け単語列)
ki を含む単語列を受ける単語列がある場合、ki はその単語列内の主辞単語から二単語ws とws+1 のみに依存すると仮定する(図１３参照)。
【数７】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_s,w_{s+1})$$

【００７０】
[ＫＭ５]
係り単語列を最大二単語列考慮(係り二単語列)
ki を含む単語列に係る単語列がある場合、ki は、そのうち最も文末側の単語列の末尾から二単語wl 、wl-1 と、最も文頭側の単語列の末尾から二単語wh 、wh-1のみに依存すると仮定する(図１３参照)。
【数８】

$$P(K|M,D,T) = \prod_{i=1}^{n} P(k_i|w_s,w_{s+1})$$

【００７１】

次に、形態素モデル（ＭＭ）について示す。形態素に付与するべき文法的属性がl 個あると仮定する。テキストつまり文字列が与えられたとき、その文字列が形態素であり、かつj(1≦j≦l)番目の文法的属性を持つとしたときの尤もらしさを確率値として求めるモデルを用いる。
テキストT が与えられたとき、順序付き形態素集合M が得られる確率は、各形態素mi(1≦i≦n) が独立であると仮定し、
【数９】

$$P(M|T) = \prod_{i=1}^{n} P(m_i|m_1^{i-1}, T)$$

と表す。ここで、mi は1 からl までのいずれかの文法的属性を表わす。
【００７２】
一方、係り受けモデル（ＤＭ）は、テキストTと順序付き形態素集合Mが与えられたとき、各単語列に対する係り受けの順序付き集合Dが得られる確率は、各々の係り受けd1・・・dn が独立であると仮定し、
【数１０】

$$P(M|T) = \prod_{i=1}^{n} P(m_i|m_1^{i-1}, T)$$

と表わす。
【００７３】
例えば、「palace troops fierce battle engage」の５つの特徴的単語訳語から「Theyare engaged in fierce battles with troops at the palace」と「I engaged thepalace's troops in fierce battles.」の２つの候補が生成されたとする。係り受けモデルにより、このうち尤もらしい係り受け構造を持つ候補が優先される。
【００７４】
以上に示すような各モデルを用い、本発明では評価部（７４）においてテキスト候補（３５'など）に評価付けを行う。
そして、評価値が最大あるいは閾値を超えるテキスト候補、あるいは評価値の上位Ｎ個を表層文に変換して出力する。
【００７５】
出力部（８０）における出力方法としては、モニタによる表示の他、音声合成を用いた発声、翻訳システムなど他の言語処理システムへのデータ出力などが可能である。また、ネットワーク接続された他のコンピュータなどにテキストデータを送出してもよい。
【００７６】
以上に示した実施例では、特徴的単語訳語の前後に語句を付加する構成を主としているが、本発明の実施においては特徴的単語訳語（主辞単語に相当するもの）そのものを補完する構成をとることもできる。前述の通り、「engage」を補完するときに、語句を付加して「They are engaged」と補完する構成を示したが、新たに「They」を補完することができる。
【００７７】
具体的には、図７の構成に図１４の要素を追加する。すなわち、特徴的単語訳語（３４）を係り受け関係語抽出部（７９）にも入力し、該部（７９）ではコーパス（７５）から該特徴的単語訳語（３４）と係り受け関係にある単語を抽出する。

そして、単語を新たな特徴的単語訳語として加え、もともと入力された特徴的単語訳語（３４）と合わせて単語列候補生成部（７２）における処理を行う。

【００７８】

例えば、「They are engaged at the palace」そのものがコーパス（７５）に無くとも、「theyare engaged」と「they are at the palace」という係り受け関係がそれぞれコーパス（７５）にあれば、それらに共通する単語「they」を新たに特徴的単語訳語として追加することによって、単語列候補生成部（７２）によって「Theyare engaged at the palace」が生成できるようになる。

【００７９】

本構成は、計量が少なく高速な特徴的単語訳語の追加が可能であるが、本発明では必ずしもコーパスから係り受け関係にある単語を抽出することに限らず、任意の特徴的単語訳語の候補を追加し、その中から評価部（７４）における評価が結果的に最も高くなるものを出力してもよい。

これによって、特徴的単語訳語にテキストの意味を決定する重要な単語が、翻訳元言語の表現特性などにより欠落していたとしても、有意な翻訳先言語のテキストが出力できるようになる。

【００８０】

本発明の機械翻訳方法において、さらに高精度な翻訳を実現する方法として、翻訳元言語テキスト（日本語テキスト（３１））の単語の依存関係を、生成にも利用することが考えられる。以下、別実施例として説述する。

【００８１】

具体的には図３の機械翻訳装置における特徴的単語抽出部（５０）を、図８に示す依存関係を解析し、それをテーブルとして記録可能な特徴的単語抽出部（５０'）に置き換える。同様に、訳語選択部（６０）は特徴的単語訳語の依存関係に変換可能な図９に示す訳語選択部（６０'）に、テキスト生成部（７０）は単語訳語の依存関係を用いてテキスト生成が可能な図１１又は図１２のテキスト生成部（７０'）（７０''）にそれぞれ置き換える。ここに述べた新たな機能以外の構成は上記実施例と同様であって、ここでは省略する。

【００８２】

図８において、「官邸前などでドゥダエフ政権部隊と激しい市街戦を展開している。」という日本語テキスト（３１）を入力する。依存関係解析部（５７）では、公知の依存関係（係り受け関係）の解析方法を用いて、図１５のように「官邸前などで」（８０）「ドゥダエフ政権部隊と」（８１）「激しい」（８２）「市街戦を」（８３）「展開している」（８４）が依存関係（８５）を有していることがわかる。ここで一般に依存関係解析部（５７）における構文解析処理にはあらかじめ形態素解析を行うことが必要であり、形態素解析部（５１）における解析結果を用いる。

【００８３】

前述の処理により特徴語決定部（５５）で決定された特徴的単語間の依存関係のみをとると、単語間依存関係抽出部（５８）では、図１６のように「官邸」（８６）「部隊」（８７）「激しい」（８９）「戦」（９０）の依存関係（９１）が抽出される。

抽出された特徴的単語間の依存関係は単語間依存関係テーブル（３６）として外部記憶装置やメモリなどに記録する。

【００８４】

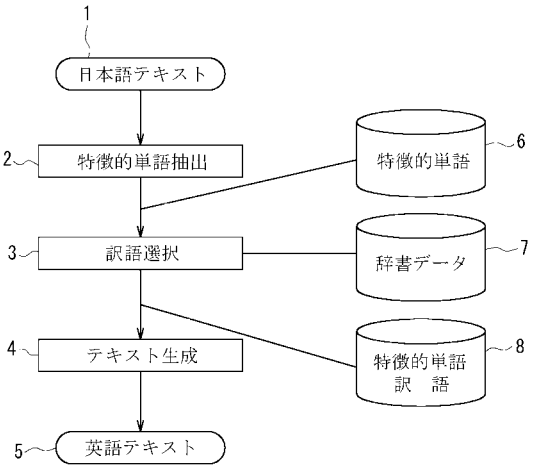
次に、図９の訳語選択部（６０'）では特徴的単語（３３）と特徴的単語訳語（３４）の関係から、日本語の単語間依存関係テーブル（３６）を、英語の単語訳語間依存関係テーブル（３７）に変換する。

具体的には、依存関係変換部（６５）を設け、対訳辞書データベース（６３）で一義的に決定できる単語、例えば部隊とtroopsのように対訳関係が明確な単語はそのまま依存関係テーブルの語を置き換える。一方、機械学習モデルを用いて訳語を選択したもの、例えば展開とengageなどのように多義性が生じるものは機械学習モデルによる類似性判定部（６

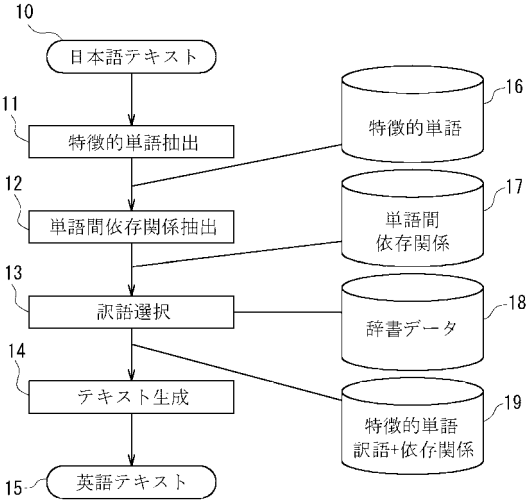
２）の判定結果を用いて単語訳語間依存関係テーブル（３７）に記録する。
【００８５】
これにより生成された単語訳語間依存関係テーブル（３７）は図１７のように、「palace」（９２）「troops」（９３）「fierce」（９４）「battle」（９５）「engage」（９６）に置き換えられた上で、依存関係（９７）が記録される。
ここで、固有名詞も特徴的単語とすると、図１５でドゥダエフが抽出され、図１８のように「Dudayev」（９８）も追加することができる。
【００８６】
図１０のテキスト生成部（７０'）ではテキスト候補生成部（７３）に抽出した単語訳語間依存関係テーブル（３７）を入力して、上述の係り受け関係の仮定を行わず、係り受け関係を決定する。
もちろん、単語訳語間依存関係テーブル（３７）に依存関係の情報がない訳語間についてはテキスト候補生成部（７３）において係り受け関係の仮定を行うこともできる。
【００８７】
本構成によると、テキスト候補が高精度に生成できるため、評価部（７４）では単語列候補生成部（７２）・テキスト候補生成部（７３）の生成結果をより厳密に評価することができる。
【００８８】
一方、図１１のテキスト生成部（７０''）では評価部（７４）に抽出した単語訳語間依存関係テーブル（３７）を入力して、テキスト候補の評価に用いることができる。
ここで、評価部（７４）は上述の言語モデル（７８）により、各単語列に対する係り受けの順序に係る確率を用いるが、単語訳語間で単語訳語間依存関係テーブル（３７）に記録された係り受け関係については、確率値を最大に設定し、当該テキスト候補の評価に用いる。
【００８９】
本構成によると、仮に単語訳語間の依存関係の一部に解析誤差が生じた場合にも、テキスト候補生成部（７３）ではそれにとらわれない候補の生成が可能なため、著しく不自然な係り受け関係は評価部（７４）において相対的に低い評価とすることができ、本発明の特徴である自然な英語テキスト（３２）の生成に寄与する。
【００９０】
本発明は以上の構成により実現されるものであるが、特徴的単語の抽出処理、訳語の選択処理、いくつかのキーワードからテキストを生成する処理はいずれも公知のあるいは今後提供される言語処理技術を用いることができる。そして、各処理の高精度化に伴ってさらに翻訳精度の向上が期待されるものである。
【００９１】
上記では説明の便宜のために、各部（４０）（５０）（６０）（７０）（８０）を別個に説述したが、これらは一体的に例えば１台のパーソナルコンピュータによって提供することができる。特に、ＣＰＵ、メモリ、入出力装置、ネットワークに接続するためのネットワークアダプタ（図示していない）、外部記憶装置などは共用することが望ましく、装置の簡略化に寄与することができる。
【００９２】
外部記憶装置に記録される文書データベース（５６）、対訳辞書・対訳コーパスデータベース（６３）、コーパス（７５）はいずれも同一のデータベースの一部又は全部を用いることが可能である。
また、これらは外部記憶装置上に記録される場合にとどまらず、ネットワーク上の複数のサーバーに記録されたものを収集するように構成してもよい。
【図面の簡単な説明】
【００９３】
【図１】本発明による機械翻訳方法のフローチャートである。
【図２】本発明による機械翻訳方法（別実施例）のフローチャートである。

【図３】本発明による機械翻訳装置の全体構成図である。
【図４】本発明における入力部の構成図である。
【図５】本発明における特徴的単語抽出部の構成図である。
【図６】本発明における訳語選択部の構成図である。
【図７】本発明におけるテキスト生成部の構成図である。
【図８】本発明における特徴的単語抽出部（別実施例）の構成図である。
【図９】本発明における訳語選択部（別実施例）の構成図である。
【図１０】本発明におけるテキスト生成部（別実施例）の構成図である。
【図１１】本発明におけるテキスト生成部（別実施例）の構成図である。
【図１２】特徴的単語訳語からのテキスト生成の例を示す説明図である。
【図１３】特徴的単語訳語と単語列との関係を示す説明図である。
【図１４】本発明に係る係り受け関係語抽出部の構成図である。
【図１５】依存関係解析部における依存関係の解析結果を示す構造木である。
【図１６】単語間依存関係抽出部における単語間依存関係テーブルの内容である。
【図１７】依存関係変換部において変換された単語間依存関係テーブルの内容である。
【図１８】同、固有名詞を特徴的単語とした時の単語間依存関係テーブルの内容である。
【符号の説明】
【００９４】
３０ 機械翻訳装置
３１ 日本語テキスト
３２ 英語テキスト
３３ 特徴的単語
３４ 特徴的単語訳語
３５ 生成テキスト
４０ 入力受理部
５０ 特徴的単語抽出部
６０ 訳語選択部
７０ テキスト生成部
８０ 出力部

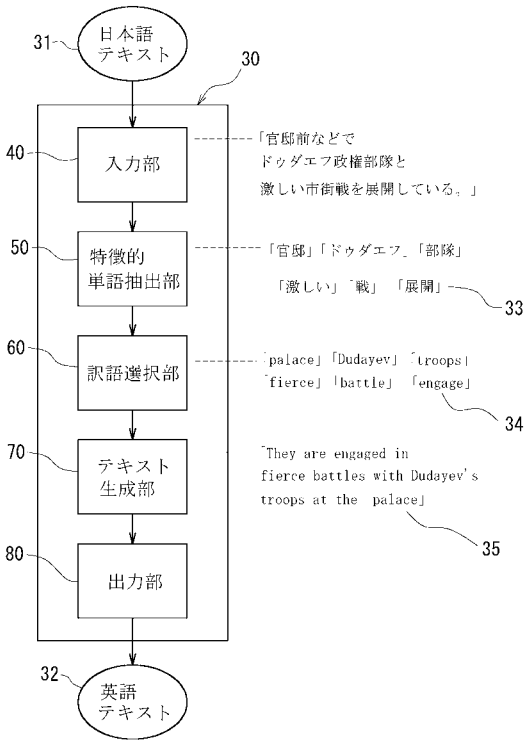
【図１】


1
日本語テキスト
2
特徴的単語抽出
3
訳語選択
4
テキスト生成
5
英語テキスト
6
特徴的単語
7
辞書データ
8
特徴的単語
訳　語

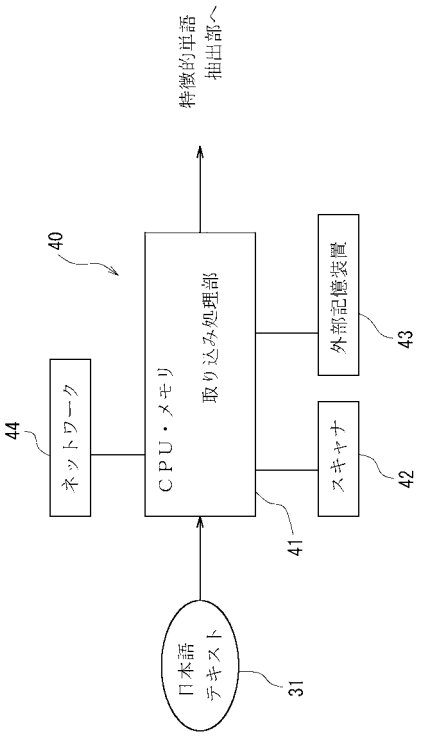
【図２】


10
日本語テキスト
11
特徴的単語抽出
12
単語間依存関係抽出
13
訳語選択
14
テキスト生成
15
英語テキスト
16
特徴的単語
17
単語間
依存関係
18
辞書データ
19
特徴的単語
訳語+依存関係

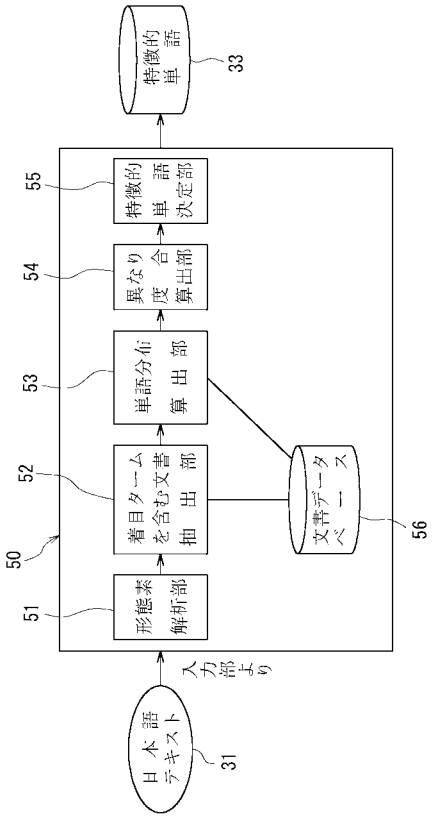
【図３】


31
日本語
テキスト
30
40
入力部
「官邸前などで
ドゥダエフ政権部隊と
激しい市街戦を展開している。」
50
特徴的
単語抽出部
「官邸」「ドゥダエフ」「部隊」
「激しい」「戦」「展開」
33
60
訳語選択部
「palace」「Dudayev」「troops」
「fierce」「battle」「engage」
34
70
テキスト
生成部
「They are engaged in
fierce battles with Dudayev's
troops at the palace」
35
80
出力部
32
英語
テキスト

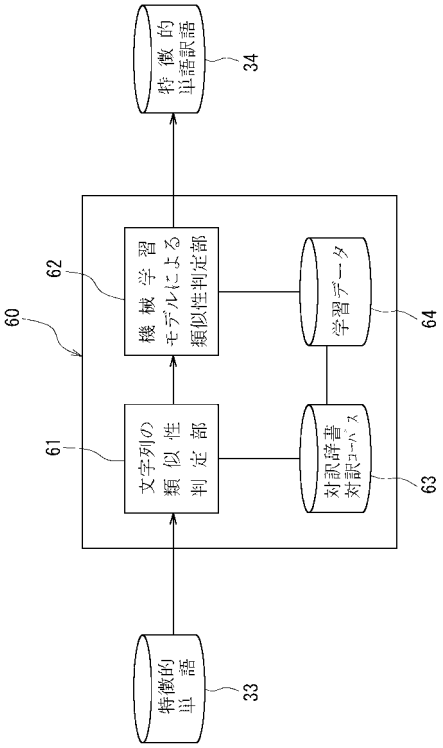
【図４】


特徴的単語
抽出部へ
40
CPU・メモリ
取り込み処理部
ネットワーク
44
外部記憶装置
43
スキャナ
42
41
日本語
テキスト
31

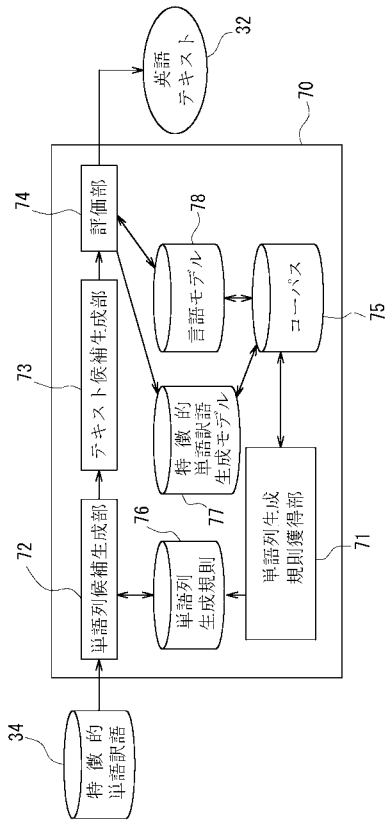
【図５】


日本語テキスト
31
入力部より
50
51
形態素解析部
52
着目タームを含む文書抽出部
53
単語分布算出部
54
異なり度合算出部
55
特徴的単語決定部
33
特徴的単語
文書データベース
56

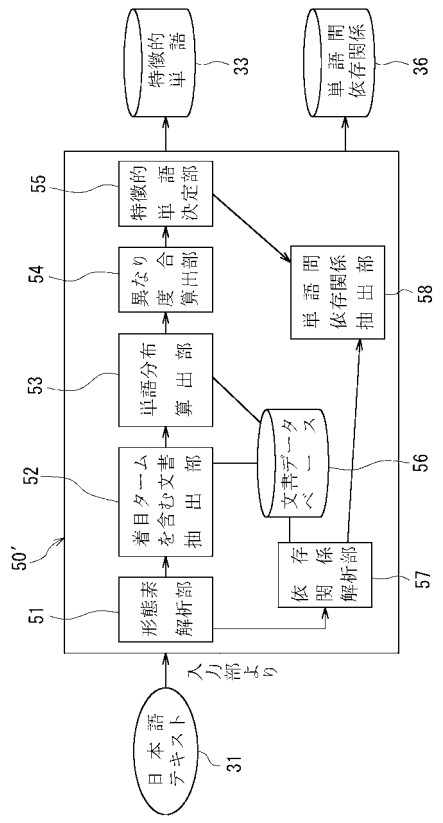
【図６】


特徴的単語
33
60
61
文字列の類似性判定部
62
機械学習モデルによる類似性判定部
63
対訳辞書 対訳コーパス
64
学習データ
特徴的単語訳語
34

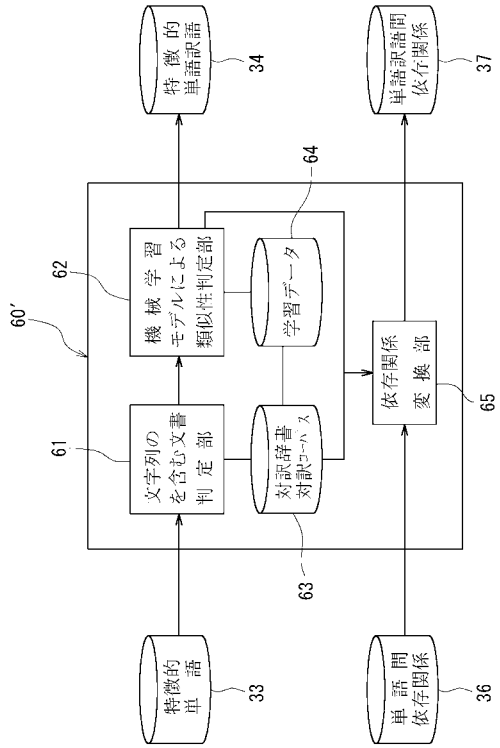
【図７】


特徴的単語訳語
34
70
71
単語列生成規則獲得部
72
単語列候補生成部
73
テキスト候補生成部
74
評価部
75
コーパス
76
単語列生成規則
77
特徴的単語訳生成モデル
78
言語モデル
英語テキスト
32

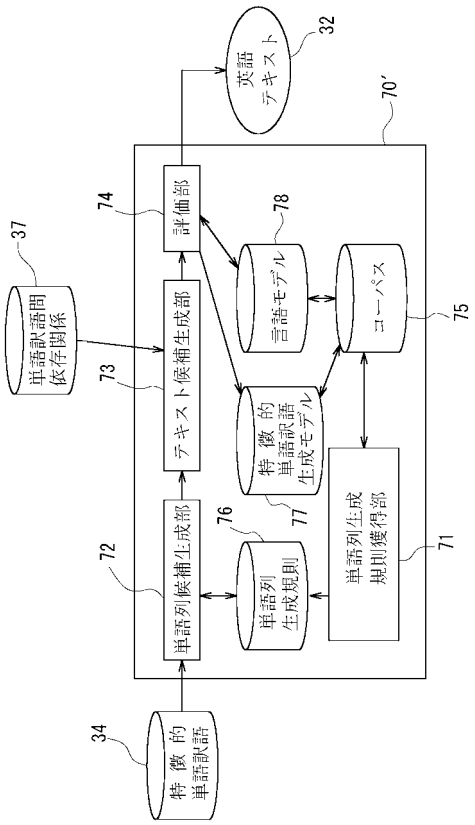
【図８】


日本語テキスト
31
入力部より
50′
51
形態素解析部
52
着目タームを含む文書抽出部
53
単語分布算出部
54
異なり度合算出部
55
特徴的単語決定部
特徴的単語
33
依存関係解析部
57
単語間依存関係抽出部
58
文書データベース
56
単語間依存関係
36

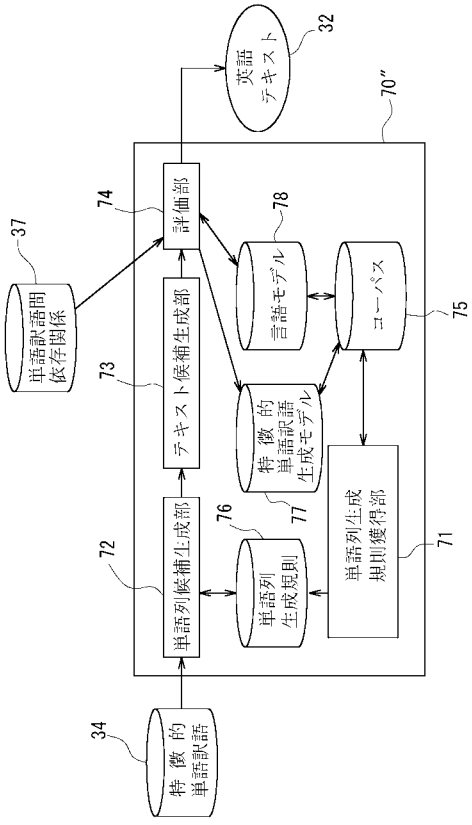
【図９】



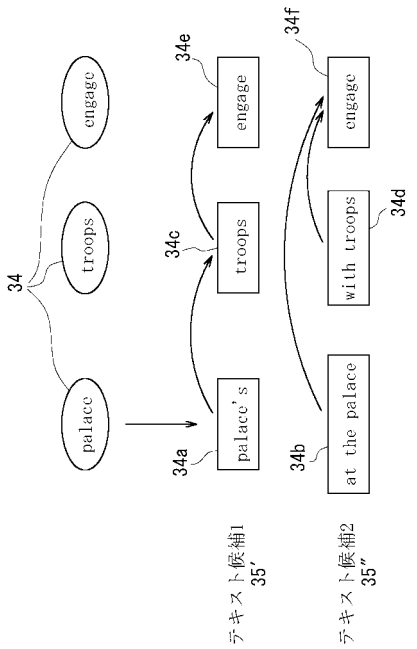
【図１０】



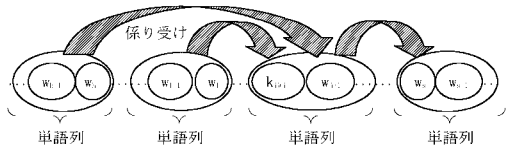
【図１１】



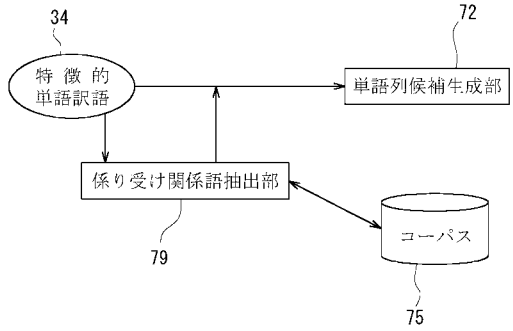
【図１２】



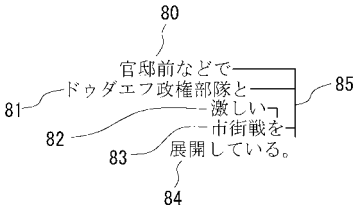
【図１３】



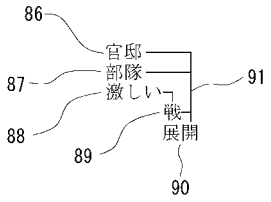
【図１４】



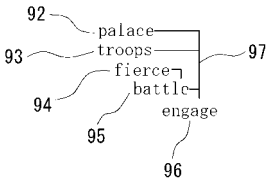
【図１５】



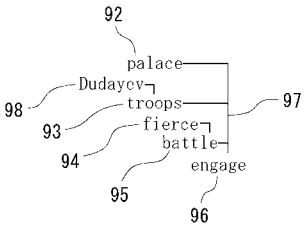
【図１６】



【図１７】



【図１８】



フロントページの続き

(56)参考文献 特公平０８－０３３８９５（ＪＰ，Ｂ２）
久光徹他，タームのｒｅｐｒｅｓｅｎｔａｔｉｖｅｎｅｓｓを測る，情報処理学会研究報告９９－ＮＬ－１３３，日本，社団法人情報処理学会，１９９９年 9月10日，Ｖｏｌ．９９，Ｎｏ．７３，ｐ．１１５－ｐ．１２２

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ １７／２７－１７／２８