(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2003－141114

（Ｐ２００３－１４１１１４Ａ）

(43)公開日　平成15年５月16日(2003.5.16)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ (参考)
G06F 17/28		G06F 17/28	P 5B091

審査請求　有　請求項の数13　ＯＬ　（全11頁）

(21)出願番号　特願2002－232922(Ｐ2002－232922)

(22)出願日　平成14年８月９日(2002.8.9)

(31)優先権主張番号　特願2001－243118(P2001－243118)

(32)優先日　平成13年８月10日(2001.8.10)

(33)優先権主張国　日本（ＪＰ）

(71)出願人　301022471
独立行政法人通信総合研究所
東京都小金井市貫井北町４－２－１

(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

(74)代理人　100090893
弁理士　渡邊　敏

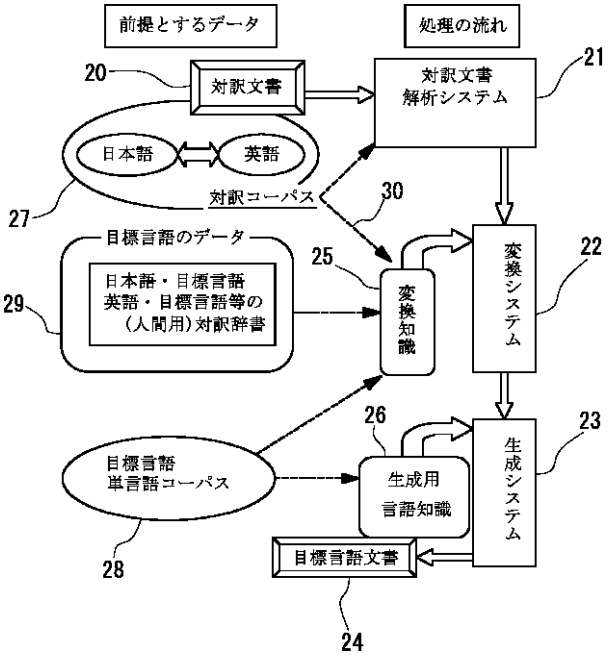
Ｆターム(参考)　5B091 CA05 CA12 CA21

(54)【発明の名称】複数言語対訳テキスト入力による第３言語テキスト生成アルゴリズム及び装置、プログラム

(57)【要約】

【課題】　主要言語間のみならず主要言語・非主要言語間における機械翻訳に用いることができる第３言語テキストの生成技術を創出すること。同時に、従来よりも高精度にテキストを生成することのできる生成技術を提供する。

【解決手段】　複数の対訳関係を有する言語テキストを入力し、両言語の対訳コーパスを用いることで、従来の単言語入力よりも高精度な第３言語テキストを生成する技術を実現する。入力後、解析過程、変換過程、生成過程の各過程を経て、目標言語文書を出力する。目標言語文書は、固有情報を自動獲得可能なため、大規模なコーパス等を必要としないことに特徴を有する。



1

【特許請求の範囲】

【請求項１】コンピュータにおける言語処理のうち、複数の言語テキストを用いて新たな第３の言語テキストを生成するアルゴリズムであって、該アルゴリズムが、

異なる言語によって記述され、翻訳元となる第１の言語と、該第１の言語と対訳関係にある少なくとも第２の言語で記述された、２つ以上の対訳テキストを入力する入力ステップ、

各対訳テキストにつき、各言語毎に、又は各言語を任意に２つ以上組み合わせて、少なくとも係り受け解析及び意味解析を含む言語解析を行い、少なくとも依存構造及び意味表現に係る言語情報を獲得する解析ステップ、

第３言語によるテキストを生成する生成ステップの各ステップを含む構成であって、

生成ステップが、

解析ステップにおいて獲得された言語情報、又は、

解析ステップの後、該解析結果に基づき、第３言語固有の変換知識を備えて言語変換を行う変換ステップを設け、該変換ステップにおける変換結果、

の少なくともいずれかを用いて第３言語によるテキストを生成することを特徴とする第３言語テキスト生成アルゴリズム。

【請求項２】前記解析ステップが、

各対訳テキストを構成する語句・文が、いかなる対訳関係を有するかについて関連づけを行う関連づけ過程、

少なくとも前記第１の言語のテキストにつき、それぞれ予め用意された解析モジュールを用いて解析する解析過程、

関連づけの結果、第１の言語のテキストと対訳関係にある少なくとも第２の言語のテキスト中の部分を予め用意された解析モジュールを用いて解析し、該各解析結果を融合する融合過程の各過程を含む請求項１に記載の第３言語テキスト生成アルゴリズム。

【請求項３】前記解析・変換・生成ステップの少なくともいずれかにおいて、

各言語に関する辞書情報又は文法情報の少なくともいずれかを含んで構成される規則的情報と、

コーパス等の実データからの学習結果による経験的情報とを用いる請求項１又は２に記載の第３言語テキスト生成アルゴリズム。

【請求項４】前記生成ステップにおいて、

第３言語の構文構造情報、又は第３言語の単語用法情報の少なくともいずれかについての情報が、該言語の既存のコーパスから一部又は全部について自動獲得して形成され、

該自動獲得された第３言語の固有情報に基づき第３言語によるテキストを生成する請求項１ないし３に記載の第３言語テキスト生成アルゴリズム。

【請求項５】言語処理のうち、複数の言語を用いて新たな第３の言語テキストを生成する装置であって、該装置

2

が、

異なる言語によって記述され、翻訳元となる第１の言語と、該第１の言語と対訳関係にある少なくとも第２の言語で記述された、２つ以上の対訳テキストを入力する入力手段、

各対訳テキストにつき、各言語毎に、又は各言語を任意に２つ以上組み合わせて、少なくとも係り受け解析及び意味解析を含む言語解析を行い、少なくとも依存構造及び意味表現に係る言語情報を獲得する解析手段、

第３言語によるテキストを生成する生成手段、

該生成手段によって生成された第３言語テキストを出力可能な出力手段の各手段を備える構成であって、

生成手段が、

解析手段において獲得された言語情報、又は、解析手段の解析結果に基づき、第３言語固有の変換知識を備えて言語変換を行う変換手段を備え、該変換手段における変換結果、

の少なくともいずれかを用いて第３言語によるテキストを生成することを特徴とする第３言語テキスト生成装置。

【請求項６】前記解析手段が、

各対訳テキストを構成する語句・文が、いかなる対訳関係を有するかについて関連づけを行う対訳関係関連づけ部、

少なくとも前記第１の言語のテキストを解析する、解析モジュール部、

該関連づけの結果、第１の言語のテキストと対訳関係にある少なくとも第２の言語のテキスト中の部分を予め用意された解析モジュールを用いて解析し、該各解析結果を融合する融合部を備える請求項５に記載の第３言語テキスト生成装置。

【請求項７】前記第３言語テキスト生成装置が、

各言語に関する辞書情報又は文法情報の少なくともいずれかを含んで構成される規則的情報と、

コーパス等の実データからの学習結果による経験的情報とを各々記憶する情報記憶手段を備えると共に、

前記解析手段・変換手段・生成手段の少なくともいずれかが、

該情報記憶手段によって記憶された各情報に基づいて解析処理を行う請求項５又は６に記載の第３言語テキスト生成装置。

【請求項８】前記第３言語テキスト生成装置が、

第３言語の構文構造情報、又は第３言語の単語用法情報の少なくともいずれかについての情報を、該言語の既存のコーパスから一部又は全部について自動獲得する第３言語固有情報獲得手段又は、

予め自動獲得された第３言語固有情報を保持可能な第３言語固有情報記憶手段の少なくともいずれかの手段を有し、

前記生成手段が、

該第３言語固有情報に基づき第３言語テキストを生成する請求項５ないし７に記載の第３言語テキスト生成装置。

【請求項９】前記第３言語テキスト生成装置における入力手段が、

紙片、書籍等の文書を電磁的記録に変換する文書取込変換手段によって変換作成されたコンピュータデータ、又は、

ハードディスク、光学的記憶装置等の電磁的記録装置から読み出されるコンピュータデータ、又は、

インターネット等のネットワーク上の電磁的記憶装置から取得可能なコンピュータデータの少なくともいずれかのコンピュータデータを該装置に入力可能である請求項５ないし８に記載の第３言語テキスト生成装置。

【請求項１０】コンピュータにおける言語処理のうち、複数の言語テキストを用いて新たな第３の言語テキストを生成するプログラムであって、該プログラムが、

異なる言語によって記述され、翻訳元となる第１の言語と、該第１の言語と対訳関係にある少なくとも第２の言語で記述された、２つ以上の対訳テキストをコンピュータ上の記憶装置又は入力装置から取得する入力部、

取得した各対訳テキストにつき、各言語毎に、又は各言語を任意に２つ以上組み合わせて、少なくとも係り受け解析及び意味解析を含む言語解析処理を行い、少なくとも依存構造及び意味表現に係る言語情報を、コンピュータ上の演算装置及び記憶装置を用いた演算処理により獲得する解析処理部、

第３言語によるテキストをコンピュータ上の演算装置及び記憶装置を用いた演算処理により生成する生成処理部

該生成処理部によって生成された第３言語テキストをコンピュータ上の記憶装置又は出力装置により出力する出力部の各部を含む構成であって、

生成処理部が、

解析処理部において獲得された言語情報、又は、

解析処理部の解析結果に基づき、第３言語固有の変換知識を備えて言語変換を行う変換処理部を設け、該変換処理部における変換結果、

の少なくともいずれかを用いて第３言語によるテキストを生成することを特徴とする第３言語テキスト生成プログラム。

【請求項１１】前記解析処理部が、

各対訳テキストを構成する語句・文が、いかなる対訳関係を有するかについて関連づけを行う対訳関係関連づけルーチン、

少なくとも前記第１の言語のテキストを解析する、解析ルーチン、

該関連づけの結果、第１の言語のテキストと対訳関係にある少なくとも第２の言語のテキスト中の部分を解析ルーチンを用いて解析し、該各解析結果を融合する融合ルーチンの各ルーチンを含む請求項１０に記載の第３言語テキスト生成プログラム。

【請求項１２】前記解析処理部・変換処理部・生成処理部の少なくともいずれかにおいて、

各言語に関する辞書情報又は文法情報の少なくともいずれかを含んで構成される規則的情報と、

コーパス等の実データからの学習結果による経験的情報とを用いる請求項１０又は１１に記載の第３言語テキスト生成プログラム。

【請求項１３】前記第３言語テキスト生成プログラムが、

第３言語の構文構造情報、又は第３言語の単語用法情報の少なくともいずれかについての情報を、該言語の既存のコーパスから一部又は全部について自動獲得した第３言語固有情報を読み出す第３言語固有情報読み出しルーチンを備え、

前記生成処理部が、

該第３言語固有情報に基づき第３言語テキストを生成する請求項１０ないし１２に記載の第３言語テキスト生成プログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、機械翻訳等における目標言語のテキストを高精度に生成する技術に関するものである。より詳しくは、複数の言語を入力し、それらの情報を融合することで目標言語テキスト生成の高精度化を図る技術である。

【０００２】

【従来の技術】近年、多くの情報がコンピュータ上に記録され、特にインターネットの普及に伴って、そのようなデジタルデータにアクセスする手段を持つ者と持たない者の格差、いわゆるデジタルデバイドの問題が大きくなっている。それに加えて、インターネット上に記録された多くの情報は、英語等の主要言語によるものが大半であり、それらの言語を解する者と解さない者との格差も大きな問題である。

【０００３】これまで、これら言語障壁によるデジタルデバイドの解消策として機械翻訳の研究は各所で行われ、内外の多くの企業や研究所が取り組んでいる。例えば、入力言語と出力言語の対訳を用いて、その言語間の翻訳に必要とされる知識を獲得する、コーパスを用いた機械翻訳の研究がおこなわれているが、これらは大規模な対訳データが存在する言語間でしか実現できない上に、知識を獲得するだけであるため、従来よりも高精度な機械翻訳には寄与しても、主要言語にしか用いることができない。

【０００４】このように従来研究されている技術は、主要言語間でのみ用いることができる技術が大半であり、言語障壁によるデジタルデバイドの解消には寄与しないと言わざるを得ない。インターネットをはじめとする情報技術の進展により、このような格差は急速に広がりつ

5

つあり、格差が致命的になる前に対処することが緊急の課題である。しかしながら、発展途上国には言語資源と技術を開発するコストを負担する能力は乏しく、情報産業が収益の伴わない高額の投資をするのは難しい。また、先進国においても、多くの非主要言語に個別に対応するだけの負担は不可能である。これらを解決するためには、低コストで非主要言語にも対応可能な言語処理手法の開発が求められているが、従来そのような技術開発は遅れている。

【０００５】さらに、現在の機械翻訳の精度は、広く実用に供する域には達していない。ひとつの文を見ただけでは十分に意味が取れず、前後の文脈を見てはじめて意味がわかるような文が存在するが、現在の自然言語処理技術は、このような文脈を扱う能力は不充分である。

【０００６】

【発明が解決しようとする課題】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、主要言語間のみならず主要言語・非主要言語間における機械翻訳に用いることができる第３言語テキストの生成技術を創出することである。同時に、従来よりも高精度にテキストを生成することのできる生成技術を提供する。

【０００７】

【課題を解決するための手段】本発明は、上記の課題を解決するために、次のような第３言語テキストの生成アルゴリズムを用いる。すなわち、そのもっとも核心的技術は、複数の言語テキストを用いて新たな第３の言語テキストを生成するものである。そして、本発明によるアルゴリズムでは次の各ステップを含んでいる。

（１）　異なる言語によって記述され、翻訳のベースとなる第１の言語と、その第１の言語と対訳関係にある少なくとも第２の言語によって記述された２つ以上の対訳テキストを入力する入力ステップ。

（２）　各対訳テキストにつき、各言語毎に、又は各言語を任意に２つ以上組み合わせて、少なくとも係り受け解析及び意味解析を含む言語解析を行い、少なくとも依存構造及び意味表現に係る言語情報を獲得する解析ステップ。

（３）　該変換ステップにおける変換結果に基づき第３言語によるテキストを生成する生成ステップ。

【０００８】そして、生成ステップが、解析ステップにおいて獲得された言語情報又は、解析ステップの後、該解析結果に基づき、第３言語固有の変換知識を備えて言語変換を行う変換ステップを設け、該変換ステップにおける変換結果、の少なくともいずれかを用いて第３言語によるテキストを生成する。

【０００９】さらに、上記解析ステップが、各対訳テキストを構成する語句・文が、いかなる対訳関係を有するかについて関連づけを行う関連づけ過程、少なくとも前記第１の言語のテキストにつき、それぞれ予め用意された解析モジュールを用いて解析する解析過程、関連づけの結果、第１の言語のテキストと対訳関係にある少なくとも第２の言語テキスト中の部分を予め用意された解析モジュールを用いて解析し、該各解析結果を融合する融合過程の各過程を含んでもよい。

6

【００１０】上記解析ステップ、変換ステップ、生成ステップの少なくともいずれかにおいて、各言語に関する辞書情報又は文法情報の少なくともいずれかを含んで構成される規則的情報と、コーパス等の実データからの学習結果による経験的情報とを用いてもよい。

【００１１】上記生成ステップにおいて、第３言語の構文構造情報、又は第３言語の単語用法情報の少なくともいずれかについての情報が、該言語の既存のコーパスから一部又は全部について自動獲得して形成され、該自動獲得された第３言語の固有情報に基づき第３言語によるテキストを生成してもよい。

【００１２】以上の方法を用いた第３言語テキストの生成装置を提供することもできる。また、以上の方法によるプログラムとして提供することも可能である。

【００１３】

【発明の実施の形態】以下、本発明の実施方法を図面に示した実施例に基づいて説明する。なお、本発明の実施形態は以下に限定されず、適宜変更可能である。本発明は、従来の機械翻訳を超えた精度で目標とする第３言語のテキスト（以下、単に目標言語と呼ぶ。）を生成するため、人手により作成された高精度の複数の対訳文書、例えば日米２カ国の言語から内容面での情報を得、対訳辞書等から変換規則を得、目標言語の文書から言語的特徴を得て、目標言語の的確な文章を生成する技術である。

【００１４】従来の自然言語処理技術は、ひとつの文を読んで、それを翻訳したり、要約したりするという通常人間が行うであろう行為を模擬するものであった。しかし一方で、計算機に文脈を扱わせる技術の確立が困難であるという致命的な欠陥があった。本発明では、例えば日本語と英語の対訳文書から情報を和や積の形で取り出し、深い意味理解を実現する。

【００１５】単に情報をこのように和で取り出すことにより情報の量を増やすという試みは他の情報処理の手法ではあったが、本発明のように対訳を用いて積極的に文の曖昧性を解消するという手法は全く新規なものであり、そこに本発明の最も大きな特徴がある。また、その理解結果を元に、目標言語の単一言語コーパスから各言語固有の情報を得て、表層の文章を生成する点も全く新規な技術である。

【００１６】図１に従来から行われている単言語文書を目標言語に変換、生成するフローチャートを、図２に本発明に係る日米対訳文書から目標言語に変換、生成するフローチャートを示す。従来の方法において、単言語文書（１０）を目標言語文書（１４）に翻訳するプロセス

は、大きく分類して、解析システム（１１）、変換システム（１２）、生成システム（１３）を経て行うのが一般的であった。それら各システム（１１）（１２）（１３）の開発に当たっては、人手による規則の作成（１５）が不可欠であって、高精度なシステム開発には大規模な文書の解析作業が必要であった。たとえば、学習に用いる大規模なテキストコーパスは莫大なコストと、研究が必要であり、現状では主要言語のみにようやく整備されつつあるものの、非主要言語において用意される望みは極めて薄い。

【００１７】そこで、本発明は、図２に示すように、主要言語等のコーパスが整備された少なくとも２つの言語を用い、それらを解析システム（２１）、変換システム（２２）、生成システム（２３）の各システムを経て目標言語文書（２４）を生成する。すなわち、第３言語テキスト生成装置は、図３に示す２つ以上の対訳テキストを入力する入力手段によって文書の入力を行う。テキストは、スキャナ（３１）から画像データとしてインターフェース（３２）を介してＣＰＵ（３３）に入力して公知のＯＣＲ処理をＣＰＵ（３３）で実行し、テキストデータに変換した上でハードディスク（３４）・メモリ（３５）のいずれかに記憶させてもよい。また、ハードディスク（３４）にあらかじめ記憶したテキストデータを読み出して入力してもよい。その他、コンピュータに備えたキーボード（３６）から対訳テキストを入力してもよいし、ネットワークを介して接続された他のコンピュータ（３７）から取得する構成でもよい。これらとＣＰＵ（３３）を接続するインターフェースには対応するＩ／Ｏデバイス、ネットワークアダプタなどを用いることができる。

【００１８】各対訳テキストにつき、各言語毎に、又は各言語を任意に２つ以上組み合わせて、言語情報の解析を行う解析手段として対訳文書解析システム（２１）に至る。さらに、少なくとも該解析ステップにおける解析結果に基づき、第３言語への言語変換を行う変換手段として変換システム（２２）、該変換ステップにおける変換結果に基づき第３言語によるテキストを生成する生成手段として生成システム（２３）を有する。これらは、別に配設する出力手段（図示しない）によって出力可能である。出力手段としては、画面表示をするモニタや、ハードディスクなどの記憶装置、ネットワーク上の他のコンピュータに対して出力を行うことができる。

【００１９】入力する言語は、例えば日本語と英語の対訳関係にある文書である。本発明では、翻訳のベースとなる第１の言語を決め、それと対訳関係にある第２の言語と共に入力する。また、入力する言語は２つ以上であればよく、例えば３言語（日本語、英語、フランス語など）によってより高精度な解析を実現することもできる。

【００２０】従来の機械翻訳システムの精度が上がらない大きな理由の一つは、言語解析の困難さによる。解析の困難さというのは、曖昧性の解消ができないということであるが、対訳を用いることにより、解析が可能になる場合がある。たとえば、日本語だけを見ていては、あるものが複数であるかどうかはわからないが、英語を見れば、その語が単数形か複数形かで判断できる。一方、英語ではその語の意味的役割がわからないが、日本語では助詞がついているので、たとえば「場所」をあらわす情報であるということがわかったりする。これは日本語と英語のように言語の体系が大きく異なる言語同士を利用することで、特に有効となる。

【００２１】従って、本発明における対訳文書の言語の組み合わせとしては、日本語と英語や、日本語と中国語、或いはその３言語を用いるなど、言語体系が異なる言語を用いると特に好適である。逆に、英語とフランス語のみ等では本発明による効果は必ずしも大きくないが、英語・フランス語・日本語のように組み合わせると、英語・日本語のみの場合よりも高精度な生成が行える可能性が高く、そのような構成でもよい。

【００２２】次に、本発明に係る解析システム（２１）につき詳述する。解析システムの構成を図４に示す。本システム（２１）は、上記入力手段でハードディスク（３４）に記憶された日英二言語の対訳文書（２０）の入力を前提に、語と語(あるいは日本語の文節のようにもう少し大きい単位)の間の依存関係をＣＰＵ（３３）で解析処理する。ＣＰＵ（３３）は必要に応じてメモリ（３５）などのコンピュータにおける諸装置・部材と連携動作する。

【００２３】本実施例では、まず入力された対訳文書（２０）について、対訳文書を構成する各文間での対訳関係の関連づけを行い、次の解析処理における解析結果の融合に用いる。すなわち、日英対訳文書（２０）は、仮に全部が逐次対訳関係にあったとしても、それらは各言語の特性、読みやすさなどにより文の数が変わるため、機械的に対訳関係を見いだすことはできない場合がある。そこで、各対訳文書（２０）を構成する文が、いかなる対訳関係を有するかについて関連づけを行う対訳関係関連づけ部（４２）の処理を行い、対訳関係にあるテキストの関連づけを行っておく。関連づけのデータは例えば日本語テキストの第１０文は英語テキストの第１１文と対訳関係にある、というように例えば日本語テキスト中にタグ付けし、ハードディスク（３４）などに記憶する。関連づけの方法には２つのテキスト間の相互関係を抽出する周知の言語処理技術を用いることができるが、例えば言語横断検索により実現することもできる。

【００２４】そして、ＣＰＵ（３３）において、係り受け解析（４０）及び意味解析（４１）の各処理を少なくとも行う。これら各解析については、すでに公知であり任意の方法を用いることができるが、例えば、すでに本件出願人らが提案している日本語の係り受けモデル（内

元清貴、村田真樹、関根聡、井佐原均、「後方文脈を考慮した係り受けモデル」、自然言語処理，Vol.7，No.5，pp.3-17（2000)、に記載）を日本語及び英語に適用することによって決定する。このモデルは、二つの語(あるいは文節)が依存関係にあるかないかを学習するもので、機械学習モデルを用いて実現される。依存関係は学習されたモデルによって計算される確率の積が一文全体で最も高くなるように決定する。

【００２５】係り受け解析（４０）において、まずベースとなる日本語テキストについて構成される文を順次解析を行うが、その際、当該文にタグ付けがされ、英語の対訳文がある場合には、当該文の係り受け解析（４０）を合わせて行い、融合処理部（４３）において両者の文で上記確率の積が最も高いものをその文の係り受け解析結果とする。これにより、日本語テキストだけを入力するよりも、他の言語の解析結果を融合して最も確率の高い結果が得られるため、格段の解析結果の向上を図ることができる。

【００２６】さらに、この依存関係構造から格解析(意味解析)を行う。依存関係の処理においては、二言語対訳入力の有効性は、依存構造における係り受けの正解率の向上で計量可能である。ここでも、上記同様に日本語テキストからの解析結果と共に、対訳関係にある文が英語テキスト中に含まれるときには、融合処理部（４３）において両者の解析結果を比較し、より確率の高い意味解析の結果を用いる。このように本発明は、解析結果において単に確率の高い方を採ることができるので、より多くの言語を入力することで容易に解析精度の向上が図られる。

【００２７】係り受け解析（４０）や意味解析（４１）については、本件出願人による特願２００１－１３９５６３号にも開示されており、意味解析（４１）の一例として固有表現の抽出処理を詳述している。固有表現の抽出は、正確な訳語選択において重要な意味解析の１つであり、第３言語への翻訳に極めて有効である。もっとも、本発明はこれまでに提案されていなかった２つ以上の対訳文書を入力し、解析、変換、生成の枠組みにおける第３言語テキストの生成を図るものであるから、解析方法は問わず、例えば周知の形態素解析を行い、そのときに対訳文書からの解析結果を融合してもよく、その融合方法も解析方法によって異なるので、任意に決めることができる。

【００２８】以上の係り受け解析・意味解析結果は、ハードディスク（３４）に記憶される。このように、解析システム（２１）の構成要素としては、それぞれの言語の係り受け解析（４０）・意味解析（４１）処理を少なくとも行う解析モジュール（４５）を備え、さらに高精度な解析のために、対訳関係関連づけ部（４２）、融合処理部（４３）の各処理を行う。

【００２９】また、本発明の解析モジュール（４５）では辞書や文法など、あらかじめ作成された規則に基づく解析を行う一方、対訳関係を関連づけ、その解析結果を融合することにより実データに基づいた解析を可能にしている。このように前者の解析による規則的情報と、後者の解析による経験的情報を融合することにより、本発明ではより高精度な解析システム（２１）の実現に寄与している。

【００３０】次に、変換システム（２２）について以下に説述する。図５に変換システムの構成を示す。前述した通り、コンピュータを用いてある言語の情報を別の言語に変換するためにはコンピュータ処理に適した言語情報が必要である。これらを人手で作ることは二言語を理解する専門家による膨大な作業を必要とするため主要言語対以外で行うことは現実的ではない。また、大量の対訳コーパスからこれらの言語情報を自動獲得する手法もあるが、今まで述べたように主要言語対以外では大量の対訳コーパスを前提とすることはできない。

【００３１】そこで、本発明では翻訳元である二言語の対訳コーパス（２７）と翻訳先言語、すなわち目標言語（ここでは例えばタイ語とする）の単言語コーパス（２８）、および、翻訳元言語と翻訳先言語との間の小規模な対訳辞書等、例えば日タイ、英タイ辞書の小規模データ（２９）を組み合わせることによって、言語情報の獲得を図る。単言語コーパス（２８）の規模は小規模でもよく、言語処理のための十分な研究、解析が期待できない言語に対しても効果的に対応できる。これによって獲得された情報が、変換知識（２５）及び生成用言語知識（２６）であり、本発明に係る変換システム（２２）は該変換知識（２５）に基づき言語間の変換を司る。

【００３２】本発明では、大規模な第３言語のコーパスを用いなくとも高精度な出力を行うために、入力する対訳コーパス（２７）と第３言語の単言語コーパス（２８）とを比較し、第３言語における固有の言語情報を自動的に獲得し、変換知識データベース（５４）を生成する。例えば、複合語句などの場合、単に各単語を辞書に基づいて変換したのでは自然な言い回しとならないことが多い。特に訳語の選択や並び順などは第３言語の固有の情報であり、それを変換知識として備えておくのが好ましい。

【００３３】そこで、本発明における変換システム（２２）では、日英タイ語句間対応部（５１）を設け、日英対訳コーパス（２７）・対訳文書（２０）と、タイ語コーパス（２８）との間で、例えば同義の語句を抽出し、それを変換知識生成部（５２）において変換知識データベース（５４）に記憶する。例えば、翻訳元言語のコーパスが日英の対訳であることから、双方の言語の対訳関係にある語句に共通して最も対応する第３言語の語句を統計的に決定してもよい。

【００３４】変換知識は上記にかぎらず、日英対訳コーパス（２７）に多く見られる構文構造と、タイ語コーパ

スに多く見られる構文構造とを統計的に対応づけし、変換知識として備えることもできる。これにより解析システム（２１）の解析結果を、タイ語固有の構文構造に変換することが可能となる。

【００３５】さらに、変換部（５３）では、その時に記憶された変換知識や、以前の翻訳によって生成された変換知識を変換知識データベース（５４）から読み出し、上記解析システム（２１）でハードディスク（３４）に記憶された依存構造及び意味表現に係る言語情報を変換する。変換方法は、単に単語の係り受け関係や、固有表現を、第３言語の変換知識に合わせて上書き修正するだけで足りる。変換された情報は再びハードディスク（３４）に記憶される。

【００３６】最後に、生成システム（２３）につき詳述する。図６に生成システムの構成を示す。生成に関する技術開発は、従来あまり系統だって行われてこなかったが、作成した文書を人間が直接読む場合、その精度は人間の「読もうとする意欲」に直結する。そこで、本発明では、生成システム（２３）も極めて重要な言語処理システムの要素として捉え、次のような技術を用いている。

【００３７】すなわち、単言語コーパス（２８）から単語の用法に関する情報を得る技術と、構文構造に関する情報を得る。二言語以上の情報を用いて理解された結果を第３言語のテキストにする場合には、当然その言語についての知識が必要となる。生成される文章の質の向上のためには、その言語固有の情報も得る必要がある。しかし、これをその言語の研究者の持つ言語直観によって規則化していくということは、膨大な作業であり、主要な言語以外でこのような規則を作成するということは現実的ではない。

【００３８】そこで、本発明に係る第３言語テキスト生成装置では、個別の言語についての情報は、個別の言語のデータを元に公知の技術により自動獲得する。すなわち、ＣＰＵ（３３）はメモリ（３５）と協働しながら、構文構造獲得部（６０）において、タイ語コーパス（２８）から語順に係る構文構造を自動的に獲得する。この獲得方法については、言語処理分野において、様々な公知の手法があるが、例えば、コーパスから語順（内元清貴、村田真樹、馬青、関根聡、井佐原均、「コーパスからの語順の学習」、自然言語処理，Vol.7，No.4，pp.163-180 (2000)、に記載）を用いることもできる。

【００３９】具体的には、解析システム（２１）、変換システム（２２）で得られた、語と語の依存構造から自然な並びの表層文を生成する。本実施例では、自然な並びであるかどうかを、語順モデルを適用することによって決定した。このモデルは、同じ語を修飾する複数の修飾語があるとき、修飾語間での自然な順序を学習するもので、周知の機械学習モデルを用いて実現される。自然な語順は学習されたモデルによって計算される確率の積が一文全体で最も高くなるように決定している。この時、自動獲得した情報、例えば学習モデルにおける確率値などは、生成用言語知識データベース（６４）に記憶し、次回以降の生成に用いてもよい。

【００４０】基本的な構文構造が確定した後、表層表現決定部（６１）で、文中の個々の語に対する適切な表層表現を決定する。表層表現の決定には、従来の言語処理における周知の生成方法を用いることができるが、例えば本件出願人らが従来提案した文末モダリティの決定手法を、格の表現をはじめとする他の表層表現にも拡張して用いることもできる。

【００４１】すなわち、文末の時制情報（村田真樹、馬青、内元清貴、井佐原均、「用例ベースによるテンス・アスペクト・モダリティの日英翻訳」、人工知能学会誌 Vol.16，No.1，pp.20-27 (2001)に記載）を獲得する方法は、テンス・アスペクト・モダリティの翻訳の問題に初めて用例ベースの手法を適用したものであり、対訳のデータベースから解析しているテンス・アスペクト・モダリティ表現によく似た対訳例（用例）を取り出し、そのデータベースから翻訳結果を出力する手法である。用例間の類似度の定義として、文末からの一致文字列（もしくは分類語彙表の分類番号も含めた文字列での一致）を使っているため簡易な構成が可能であり、また、他の表層表現にも容易に適用できる。

【００４２】以上により、従来ではぎこちないテキストが出力されることが多かったコンピュータ生成による文書を、コーパスに示される実際の文章での流暢さに基づいたレベルにまで向上させることが可能となる。また、単言語コーパスからの単語用法情報を自動獲得し、生成用言語知識（２６）に加えることもできる。

【００４３】以上、本発明における第３言語テキスト生成装置の解析手段、変換手段、生成手段につき説述したが、本発明の実施においては必ずしも変換手段を設けなくともよい。すなわち、本発明で言う変換手段は、出力する言語固有の変換知識を有するものであるが、明示的に変換手段を配設しなくともよい。例えば、解析手段や生成手段の有する言語情報に関する知識・情報で十分に生成まで行える場合には、変換手段として独立した手段ではなく、解析手段による解析結果から、生成手段によって直接第３言語を生成することができる。

【００４４】また、本装置では、入力手段・出力手段についても様々な形態を考えることができる。入力手段は、多様な媒体を介して流通する情報から入力することが考えられる。例えば、紙片、書籍等の文書を電磁的記録に変換可能な文書取込変換手段を有する。これはスキャナと文字認識装置・ソフトウェアによってすでに容易に実現可能であり、本発明の装置に内蔵することによって、例えば日英の２言語で記述された対訳関係にある書籍を読みとることで、タイ語等の第３言語テキストを出力する構成を実現できる。出力には、表示装置による表

示、記録装置への書き出し、インターネット等のネットワーク上への掲載等、その出力手段は任意である。

【００４５】また、ハードディスク、光学的記憶装置等の電磁的記録装置から読み出されるコンピュータデータは、より簡易に読み出し、かつ入力することが可能である。特に、近年ユニコード等の多言語に対応した文字コードが開発されており、同時に複数の言語、特に非主要言語であっても同時に扱うことが可能になってきた。このようなコードを用いることで、円滑に複数の言語を同時に扱うことが可能であり、上記電磁気的記録装置への記録、読み出しは容易である。

【００４６】さらに、本発明が大きく効果を有する用途として、インターネット等のネットワーク上のコンピュータに付設される電磁的記憶装置から取得可能なコンピュータデータを入力することが考えられる。インターネット上では、特に主要言語が用いられる地域でコンピュータの普及が進んでいることもあり、流通する情報は多くが主要言語で記述されている。また、多国籍企業のホームページ等は、すでに主要言語間の人手による高精度な翻訳がなされており、本発明の技術を用いることで、未だ翻訳が行われていない多くの非主要言語への変換が可能となる。従って、該装置の入力手段がインターネット等のネットワークに接続された電磁気的記録装置からコンピュータデータを取得し、本装置へ入力することは非常に効果的である。

【００４７】上記では本発明の一実施形態として、第３言語テキスト生成装置を挙げたが、本発明は単にコンピュータのアルゴリズムとして提供することもできるし、また、プログラムとして実現し、任意のコンピュータ上で動作させることもできる。また、本発明によって構成されたプログラムを、ネットワーク上で流通させることもできる。

【００４８】

【発明の効果】本発明は、以上の構成を備えるので、次の効果を奏する。請求項１に記載の第３言語テキスト生成アルゴリズムによると、複数の言語で書かれた同じ内容の文章を合わせて解析することにより、正確な意味理解を行い、入力とは異なる第３の言語で適切にテキストを生成することができる。また、必要に応じて変換過程を備えることにより、より高精度化に寄与する。これにより、発展途上国等への母国語による情報提供が可能になる。また、この手法が確立すれば、新しい言語への対応は、その言語に関する言語情報の獲得が主たる開発要素となり、それぞれの国でも対応できると思われる。今後とも、英語で作られた文書を人手をかけて高品位の日本語文書とすることは引き続き大量に行われるであろうが、このような文書がすべて、他のアジアの多くの言語にも高品位に翻訳されていくことは考えにくい。本発明により、タイ語などのアジアの諸言語等への翻訳水準が飛躍的に向上することが可能である。本技術の確立により、デジタルデバイドに悩む多くの開発途上国が、独自の努力と多少の支援により、そこから抜け出すことが可能となる。さらに、従来の単言語からの翻訳に比して、飛躍的に高精度な第３言語テキストの生成を低コストで可能にすることができる。

【００４９】請求項２に記載の第３言語テキスト生成アルゴリズムによると、解析ステップが関連づけ過程、解析過程、解析結果の融合過程を含み、該解析結果を変換ステップや生成ステップにおいて用いることにより、高精度な第３言語テキストの生成方法を実現することができる。

【００５０】請求項３に記載の第３言語テキスト生成アルゴリズムによると、解析ステップ・変換ステップ・生成ステップの各ステップのいずれかにおいて、規則的情報と経験的情報を融合することで、第３言語テキストの生成方法の高精度化に寄与する。

【００５１】請求項４に記載の第３言語テキスト生成アルゴリズムによると、第３言語の固有情報を自動獲得することが出来るので、大規模な第３言語のコーパスを用意することなく、多くの非主要言語に対応可能な第３言語テキスト生成方法を実現することができ、デジタルデバイド問題に効果的に対応することができる。同時に、生成方法の低コスト化にも寄与する。

【００５２】請求項５に記載の第３言語テキスト生成装置によると、複数の言語で書かれた同じ内容の文章を合わせて解析することにより、正確な意味理解を行い、入力とは異なる第３の言語で適切にテキストを生成する生成装置を提供することができる。該装置の提供によって、低コストで、多くの非主要言語で出力可能になり、ひいてはデジタルデバイドの解消につながる。

【００５３】請求項６に記載の第３言語テキスト生成装置によると、解析手段が対訳関係の関連づけ、解析、解析結果の融合を行うことによって、低コストで高精度な第３言語テキス生成装置を実現することができる。

【００５４】請求項７に記載の第３言語テキスト生成装置によると、解析手段・変換手段・生成手段の各手段のいずれかにおいて、規則的情報と経験的情報を融合することによって、第３言語テキストの生成装置の高精度化に寄与する。

【００５５】請求項８に記載の第３言語テキスト生成装置によると、第３言語の固有情報を自動獲得することが出来るので、大規模な第３言語のコーパスを用意することなく、多くの非主要言語に対応可能な第３言語テキスト生成装置を提供することができる。同時に、該装置の低コスト化にも寄与する。

【００５６】請求項９に記載の第３言語テキスト生成装置によると、様々に流通する情報から自由に情報を入力する入力手段を有するので、近年の多様な情報流通形態にも柔軟に対応することが可能である。特に、インターネット上の情報を効率よく入力し、インターネット上に

15

出力することで、非主要言語による情報提供が容易になる。

【００５７】請求項１０に記載の第３言語テキスト生成プログラムによると、複数の言語で書かれた同じ内容の文章を合わせて解析することにより、正確な意味理解を行い、入力とは異なる第３の言語で適切にテキストを生成する第３言語テキスト生成プログラムを提供することができる。それらの提供によって、低コストで、多くの非主要言語で出力可能になり、ひいてはデジタルデバイドの解消につながる。

【００５８】請求項１１に記載の第３言語テキスト生成プログラムによると、解析処理部が関連づけ、解析、解析結果の融合を行うことによって、低コストで高精度なプログラムを実現することができる。

【００５９】請求項１２に記載の第３言語テキスト生成プログラムによると、解析処理部・変換処理部・生成処理部の各処理のいずれかにおいて、規則的情報と経験的情報を融合することによって、それらの高精度化に寄与する。

【００６０】請求項１３に記載の第３言語テキスト生成プログラムによると、第３言語の固有情報を自動獲得することが出来るので、大規模な第３言語のコーパスを用意することなく、多くの非主要言語に対応可能な第３言語テキスト生成プログラムを提供することができ、デジタルデバイド問題に効果的に対応することができる。同時に、それらの低コスト化にも寄与する。

16

【図面の簡単な説明】

【図１】従来の目標言語文書生成フローチャートである。

【図２】本発明による目標言語文書生成フローチャートである。

【図３】本発明に係る第３言語テキスト生成装置の入力手段の構成図である。

【図４】本発明に係る第３言語テキスト生成装置の解析システムの構成図である。

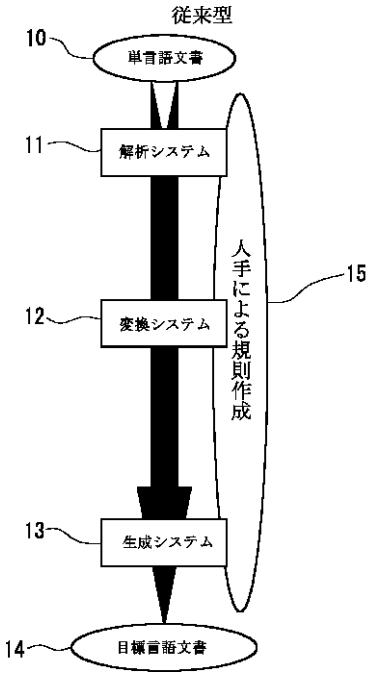
【図５】本発明に係る第３言語テキスト生成装置の変換システムの構成図である。

【図６】本発明に係る第３言語テキスト生成装置の生成システムの構成図である。

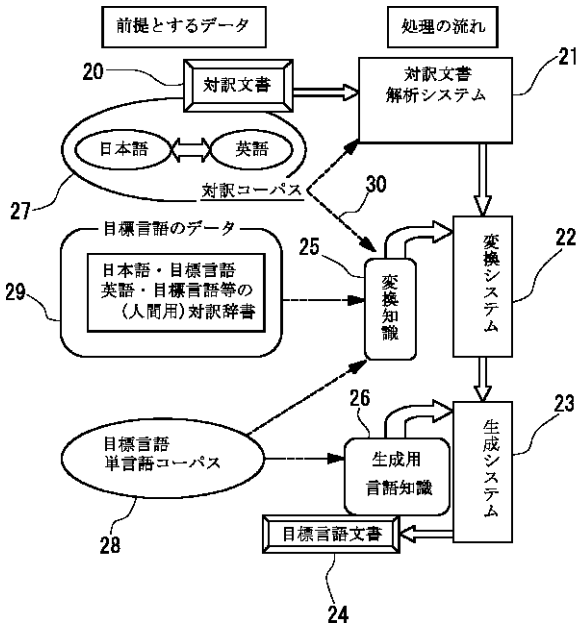
【符号の説明】

２０　対訳文書
２１　対訳文書解析システム
２２　変換システム
２３　生成システム
２４　目標言語文書
２５　変換知識
２６　生成用言語知識
２７　対訳コーパス
２８　単言語コーパス
２９　目標言語の小規模データ
３０　対訳コーパスから変換知識を取得する処理を示す矢印

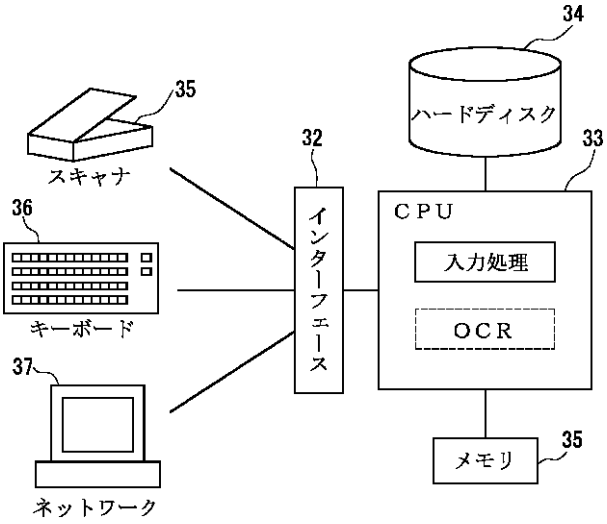
【図１】



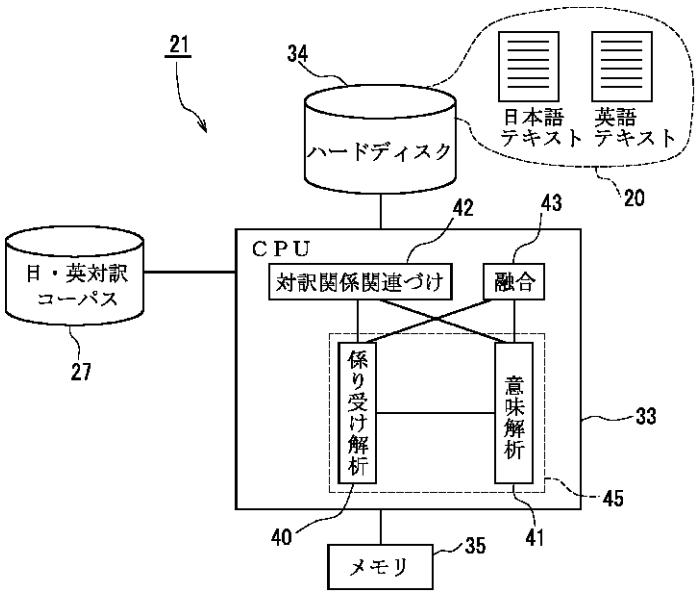
【図２】



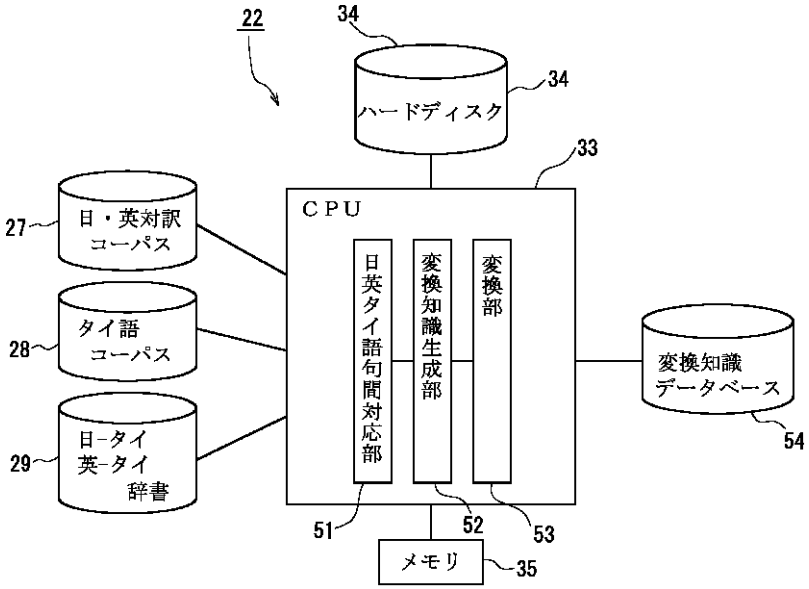
【図３】


34
ハードディスク
35
スキャナ
32
33
インターフェース
CPU
36
入力処理
キーボード
OCR
37
35
メモリ
ネットワーク

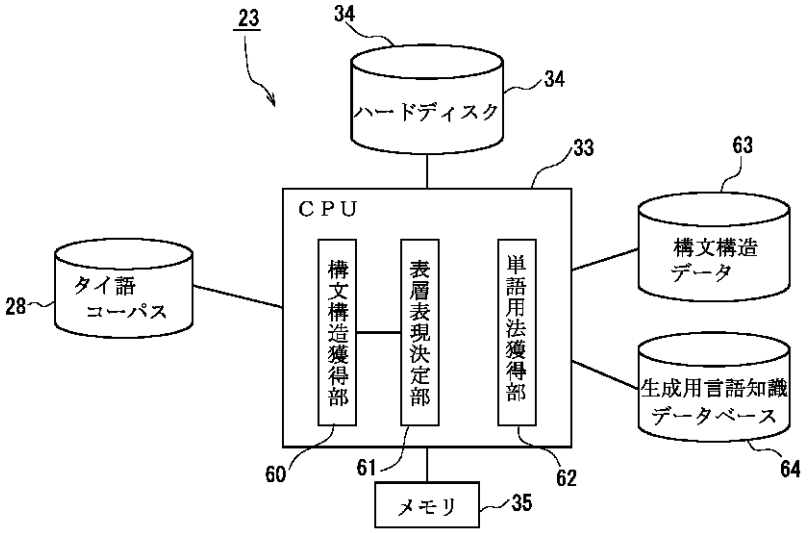
【図４】


21
34
日本語テキスト
英語テキスト
ハードディスク
20
42
43
CPU
日・英対訳コーパス
対訳関係関連づけ
融合
係り受け解析
意味解析
33
27
45
40
41
メモリ
35

【図５】


22
34
ハードディスク
34
33
CPU
日・英対訳
コーパス
27
タイ語
コーパス
28
日−タイ
英−タイ
辞書
29
日英タイ語句間対応部
変換知識生成部
変換部
変換知識
データベース
54
51
52
53
メモリ
35

【図６】


23
34
ハードディスク
34
33
63
CPU
タイ語
コーパス
28
構文構造獲得部
表層表現決定部
単語用法獲得部
構文構造
データ
生成用言語知識
データベース
64
60
61
62
メモリ
35