(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2003－141110
（Ｐ２００３－１４１１１０Ａ）
(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 17/27		G06F 17/27	Ｊ　5B091

審査請求　有　請求項の数３　ＯＬ　（全12頁）

(21)出願番号　特願2001－331458(Ｐ2001－331458)

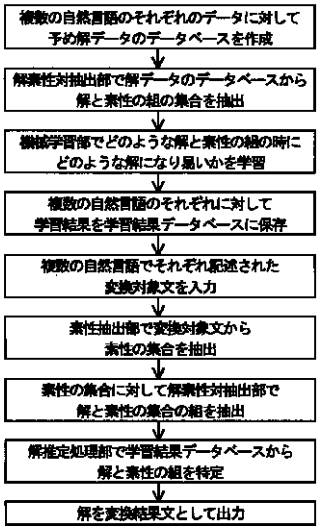
(22)出願日　平成13年10月29日(2001.10.29)

(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　村田　真樹
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
(74)代理人　100090893
弁理士　渡邊　敏

Ｆターム(参考)　5B091 AA01 AA15 BA03 CA01 EA01

(54)【発明の名称】複数言語入力での言語処理方法及び言語処理装置

(57)【要約】
【課題】　大量の変換規則を用意する必要がなく、複数の自然言語で記述された変換対象文から他の自然言語及び／又は同じ自然言語で記述された変換結果文への言語変換及び／又は複数の自然言語における言語解析を行うことができる複数言語入力での言語処理方法及び言語処理装置を提供する。
【解決手段】　複数の自然言語で記述された処理対象文から、他の自然言語及び／又は同じ自然言語で記述された処理結果文への処理を行う言語処理方法であって、前記処理を行う際に、前記他の自然言語及び／又は前記同じ自然言語への言語変換及び／又は前記同じ自然言語で記述されたどのような文又は文章になり易いかなどの言語解析を学習させる機械学習手法を用いたものである。



1

【特許請求の範囲】
【請求項１】　複数の自然言語で記述された処理対象文に対して、他の自然言語及び／又は同じ自然言語で記述された処理結果文への言語変換及び／又は特定の言語現象を明らかにする言語解析を行う言語処理方法であって、
前記言語変換を行う際に、前記他の自然言語及び／又は前記同じ自然言語で記述されたどのような文又は文章等になり易いかを学習及び／又は前記言語解析を行う際に、前記言語現象の解析においてどのような解析結果になり易いかを学習させる機械学習手法を用いたことを特徴とする複数言語入力での言語処理方法。
【請求項２】　複数の自然言語で記述された処理対象文に対して、他の自然言語及び／又は同じ自然言語で記述された変換結果文への言語変換及び／又は特定の言語現象を明らかにする言語解析を行う言語処理装置であって、
前記複数の自然言語で記述された処理対象文を入力する入力手段と、
前記入力手段により入力された前記複数の自然言語のデータに、前記他の自然言語及び／又は前記同じ自然言語への言語変換結果である解の情報が付与された解データ及び／又は特定の言語現象を明らかにする言語解析結果である解の情報が付与された解データを保存する解データベース部と、
前記解データから、前記解と解析に用いる細かい情報の１単位である素性の組を抽出する解素性対抽出部と、
前記言語変換の際に、前記解と素性の組から、どのような解になり易いかを学習及び／又は前記言語解析の際に、前記解と素性の組から、どのような解析結果になり易いかを学習する機械学習部と、
前記機械学習部で学習した結果を保存する学習結果データベース部と、
入力された複数の自然言語で記述された処理対象文から、素性を取り出す素性抽出部と、
前記素性抽出部から取り出された素性の集合に対して、前記学習結果データベース部に保存された前記学習した結果から解を特定する解推定処理部と、
を備えたことを特徴とする複数言語入力での言語処理装置。
【請求項３】　前記解素性対抽出部は、解と素性の組及び解候補と素性の組を抽出することを特徴とする請求項２記載の複数言語入力での言語処理装置。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】この発明は、ある自然言語で記述された変換対象文を、他の自然言語及び／又は同じ自然言語で記述された処理結果文に言語変換及び／又は特定の言語現象を明らかにする言語解析を行う言語処理に関し、特に、複数の自然言語で記述された処理対象文

2

を他の自然言語及び／又は同じ自然言語で記述された処理結果文に言語変換をする際及び／又は特定の言語現象を明らかにする言語解析をする際に、機械学習手法を用いる複数言語入力での言語処理方法及び言語処理装置に関するものである。
【０００２】
【従来の技術】言語処理には、各言語の形態素解析、構文解析、格解析等を行う言語解析と他の言語への翻訳を行う言語変換とがある。ある自然言語から他の自然言語言語変換する従来の言語処理技術として、機械翻訳がある。機械翻訳では、ある自然言語で記述された文又は文章等を他の自然言語で記述された文又は文章等に言語変換する。また、同一の言語間における文又は文章の言語変換、例えば、要約文を自動生成あるいは文章を推敲する言語処理技術も用いられるようになってきている。
【０００３】前記機械翻訳では、例えば、図３に示すように、ＣＰＵ（中央演算処理装置）、メモリ、データ保存装置等からなるコンピュータ本体と周辺機器とから構成される言語処理装置３０において、先ず、対象とする問題の答えである解のデータベースを作成して解データベース部３１に保存しておく。前記解のデータベースには、入力されたある自然言語のデータに前記解の情報が付与されている。
【０００４】次に、前記解データベース部３１から各事例毎に、解素性対抽出部３２で解と素性の集合の組を抽出する。前記素性は、解析に用いる情報の細かい１単位を意味し、前後の単語自体及び品詞、解析する単語自体及び品詞、解の単語及び品詞などである。
【０００５】次に、前記解と素性の集合の組から、どのような素性の時にどのような解になり易いかを機械学習部３３で学習する。この学習結果を解学習結果データベース部３４に保存する。
【０００６】ここまでは、予め準備しておく。ここから、先ず、解を求めたい文又は文章である変換対象文３５を入力する。入力された変換対象文３５から、素性抽出部３６において素性の集合を取り出し、それらを解推定処理部３７に渡す。
【０００７】解推定処理部３７では、渡された素性の集合の場合、どのような解になり易いかを前記解学習結果から特定する。最後に、特定された解を変換結果文３８として出力する。
【０００８】以上で示したように、機械翻訳では、機械学習を用い、ある自然言語で記述された文又は文章等から他の自然言語で記述されるどのような文又は文章になり易いかを特定して変換する。
【０００９】また、前記形態素解析、構文解析、格解析等を行う言語解析においても同様に、解析に用いる素性を設定し、機械学習を用いてどのような解析結果になりやすいかを学習させることが行われている。
【００１０】また、ここで示した機械学習に基づく方法

の他に人手でパターンマッチ規則を作成し、これを用いて機械翻訳、言語解析を行うこともある。現状の実運用されている機械翻訳、言語解析ではむしろ、この人手で作成した規則に基づいて行っているものの方が主である。

【００１１】また、同一自然言語間での文の言語変換処理では、一般に、変換前の語・句・文などのパターンと変換後の語・句・文などのパターンとの対からなる変換規則を大量に用意し、いわゆるパターン・マッチングによって入力文中に現れる処理前のパターンを探し出し、該当するパターンがあれば、それを処理後の語・句・文などのパターンに置き換える処理を行っている。

【００１２】

【発明が解決しようとする課題】しかしながら、前記従来の機械学習を用いた言語変換は、ある自然言語で記述された文又は文章などから他の自然言語及び／又は同じ自然言語で記述された文又は文章などへの言語変換に用いられているだけで、複数の自然言語で記述された文又は文章などから他の自然言語及び／又は同じ自然言語で記述された文又は文章などへの言語変換に用いられていなかった。また、前記従来の機械学習を用いた言語解析は、ある一つの自然言語における言語解析にしか用いられていなかった。

【００１３】従って、前記言語変換あるいは前記言語解析を行おうとすると、従来のパターン・マッチングを用いて行うしかなく、この場合、大量の変換規則を用意しなければならないという問題点があった。

【００１４】本発明は、前記従来の問題点を解決するためになされたもので、大量の変換規則を用意する必要がなく、複数の自然言語で記述された処理対象文から他の自然言語及び／又は同じ自然言語で記述された処理結果文への言語変換及び／又は複数の自然言語における言語解析を行うことができる複数言語入力での言語処理方法及び言語処理装置を提供することである。

【００１５】また、機械学習を用いる方法は、複数の情報を素性によって容易に利用できるために、本課題の複数言語入力のように複数の情報が複雑に入力される課題に対して、まさにうってつけである。

【００１６】

【課題を解決するための手段】本発明は、複数の自然言語で記述された処理対象文に対して、他の自然言語及び／又は同じ自然言語で記述された処理結果文への言語変換及び／又は特定の言語現象を明らかにする言語解析を行う言語処理方法であって、前記言語変換を行う際に、前記他の自然言語で記述されたどのような文又は文章等になり易いかを学習及び／又は前記言語解析を行う際に、前記言語現象の解析においてどのような解析結果になり易いかを学習させる機械学習手法を用いたものである。

【００１７】また、本発明は、複数の自然言語で記述された処理対象文に対して、他の自然言語及び／又は同じ自然言語で記述された処理結果文への言語変換及び／又は特定の言語現象を明らかにする言語解析を行う言語処理装置であって、前記複数の自然言語で記述された処理対象文を入力する入力手段と、前記入力手段により入力された前記複数の自然言語のデータに、前記他の自然言語及び／又は前記同じ自然言語への言語変換結果である解の情報が付与された解データ及び／又は特定の言語現象を明らかにする言語解析結果である解の情報が付与された解データを保存する解データベース部と、前記解データから、前記解と解析に用いる細かい情報の１単位である素性の組を抽出する解素性対抽出部と、前記言語変換の際に、前記解と素性の組から、どのような解になり易いかを学習及び／又は前記言語解析の際に、前記解と素性の組から、どのような解析結果になり易いかを学習する機械学習部と、前記機械学習部で学習した結果を保存する学習結果データベース部と、入力された複数の自然言語で記述された処理対象文から、素性を取り出す素性抽出部と、前記素性抽出部から取り出された素性の集合に対して、前記学習結果データベース部に保存された前記学習した結果から解を特定する解推定処理部とを備えたものである。

【００１８】また、前記解素性対抽出部は、解と素性の組及び解候補と素性の組を抽出するものでもよい。

【００１９】

【発明の実施の形態】以下、図面を参照して、本発明の実施形態について説明する。図１は、本発明による複数言語入力での言語処理装置の実施形態を示すブロック図である。

【００２０】図１に示すように、本発明の実施形態の複数言語入力での言語処理装置１０は、ＣＰＵ（中央演算処理部）、データを一時保存するメモリ、データを保存するデータ保存部、例えば、ハードディスク等を有する本体部及び表示装置であるＣＲＴ等、必要に応じた周辺機器を備えたコンピュータで構成されていて、複数の自然言語で記述された処理対象文を入力とし、これに対して他の自然言語及び／又は同じ自然言語への言語変換及び／又は特定の言語現象を明らかにする言語解析を行った結果を、処理結果文として出力する。

【００２１】言語処理装置１０は、複数の自然言語、ここでは２つの自然言語でそれぞれ記述された文又は文章などを入力する入力手段（図示せず）、例えばキーボードを備え、予め前記キーボードにより入力された前記２つの自然言語のそれぞれのデータに、他の自然言語及び／又は同じ自然言語への翻訳である言語変換及び／又は同じ自然言語への形態素解析、構文解析、格解析等の言語解析を行った処理結果である解の情報が付与された解データを保存する解データベース部１１を有する。解データベース部１１には、言語１と言語２についての解データがそれぞれ保存されている。

5

【００２２】また、前記解の情報における解は、対象とする問題の答えであり、前記言語変換の場合、変換先の言語表現である。従って、前記解の情報は、変換先の言語表現に関する情報である。また、前記言語解析の場合、形態素解析であるならば、前記解は品詞であり、前記解の情報は品詞に関する情報である。

【００２３】そして、言語処理装置１０は、解データベース部１１に保存されている解データから、前記解と解析に用いる細かい情報の１単位である素性の組（図中では、解－素性対と表す）とを抽出する解素性対抽出部（図中では、解－素性対抽出部１２と表す）と、前記解と素性の組から、前記変換の際に、どのような解になり易いかを学習する機械学習部１３と、機械学習部１３で学習した結果を保存する学習結果データベース部１４とを備えている。

【００２４】解－素性対抽出部１２では、解データベース部１１に保存されている解データを取り出し、各事例ごとに、解と素性の組を抽出すると共に、機械学習部１３で学習した結果、新たに得られた保存すべき解データを解データベース部１１に保存する。

【００２５】機械学習部１３では、言語１と言語２のそれぞれについて、処理対象文１５から処理結果文１６に処理する際に、それぞれの解と素性の組からどのような解になり易いか、即ち、どのような解と素性の組み合わせの時に解である確率が高いかを学習し、学習した結果を学習結果データベース部１４に保存する。この学習は、言語１と言語２のそれぞれに対して行い、学習結果はそれぞれ別々に分類され保存される。

【００２６】更に、言語処理装置１０は、入力された２つ自然言語の処理対象文１５である言語１と言語２から、素性を取り出す素性抽出部１７と、素性抽出部１７から取り出された素性の集合に対して、学習結果データベース部１４に保存された前記学習した結果から解を特定する解推定処理部１８とを備えている。

【００２７】処理対象文１５の言語１と言語２は、それぞれ素性抽出部１７に入力されてそれぞれの素性が取り出され、取り出されたそれぞれの素性の集合に対して、解－素性対抽出部１２から解と素性の集合の組を取り出し、それを解推定処理部１８に渡す。

【００２８】解推定処理部１８では、渡された解と素性の集合の組から、学習結果データベース部１４に保存された学習した結果に基づき解を特定する。特定された解は、出力されて処理結果文１６として得られ、必要に応じて保存される。

【００２９】尚、解－素性対抽出部１２は、解と素性の組を抽出するばかりでなく、必要に応じて、解と素性の組を抽出すると共に、解の候補となる解候補と素性の組を抽出するものでもよい。ここで、前記解の候補は、前記解以外の解の候補を意味する。

【００３０】解と素性の集合の組を正例、解候補と素性の集合の組を負例とすると、解若しくは解候補と素性の集合の組から、どのような解若しくは解候補と素性の集合の時に、正例である確率が高いかあるいは負例である確率が高いかを機械学習部１３で学習し、その結果を学習結果データベース部に保存する。

【００３１】解推定処理部１８では、解－素性対抽出部１２から抽出された解候補と素性の集合の全ての組に対して、渡された解候補と素性の組について正例、負例である確率を求め、最も正例である確率が高い解候補を解として出力し、処理結果文１６として保存する。

【００３２】次に、本発明による複数言語入力での言語処理方法の実施形態について説明する。本発明の実施形態の複数言語入力での言語処理方法は、複数の自然言語で記述された文又は文章等の処理対象文に対して、言語変換および言語解析を行う際に、前記他の自然言語及び／又は前記同じ自然言語で記述されたどのような文又は文章等になり易いか及び／又は言語解析においてどのような解析結果になり易いかを学習させる機械学習手法を用いたものである。

【００３３】即ち、図２に示すように、複数の自然言語のそれぞれのデータに、他の自然言語及び／又は同じ自然言語への翻訳である言語変換及び／又は同じ自然言語への形態素解析、構文解析、格解析等の言語解析を行った処理結果である解の情報が付与された解データのそれぞれのデータベースを作成する。

【００３４】次に、解素性対抽出部で、それぞれのデータベースから解データを取り出し、各事例ごとに、解と素性の組の集合を抽出する。

【００３５】次に、機械学習部で、どのような解と素性の組の時にどのような解になり易いかを学習する。

【００３６】次に、複数の自然言語のそれぞれについて学習した結果を学習結果データベースに保存する。

【００３７】次に、複数の自然言語で記述された処理対象文を、それぞれ素性抽出部に入力し、それぞれの処理対象文からそれぞれの素性の集合を抽出する。

【００３８】次に、それぞれの素性の集合に対して、解素性対抽出部から解と素性の集合の組を抽出し、それを解推定処理部に渡す。

【００３９】次に、解推定処理部で、渡された解と素性の集合の組から、学習結果データベースから学習した結果に基づき解を特定する。

【００４０】最後に、特定された解を処理結果文として出力する。

【００４１】尚、解素性対抽出部は、解と素性の組及び解候補と素性の組を抽出するものでもよい。ここで、前記解の候補は、前記解以外の解の候補を意味する。

【００４２】その場合には、解と素性の集合の組を正例、解候補と素性の集合の組を負例とすると、解若しくは解候補と素性の集合の組から、どのような解若しくは解候補と素性の集合の時に、正例である確率が高いかあ

るいは負例である確率が高いかを機械学習部で学習し、その結果を学習結果データベース部に保存する。

【００４３】解推定処理部では、解素性対抽出部から抽出された解候補と素性の集合の全ての組に対して、渡された解候補と素性の組について正例、負例である確率を求め、最も正例である確率が高い解候補を解として出力する。

【００４４】以上示したように、本発明の実施形態の複数言語入力での言語処理方法によれば、機械学習手法を用い、機械学習する際に、素性を複数の自然言語、例えば、２つの自然言語から取るだけで済むので、大量の変換規則を用意する従来の言語処理方法よりも処理が容易である。

【００４５】

【実施例】（実施例１）日本語の言語解析である形態素解析の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【００４６】日本語と英語の２つの自然言語のデータ（以下、対訳データという）の場合を考える。前記解データは、日本語と英語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。対訳データは、「＜きょうだい＞で待つ。」と「Ｉ ｗａｉｔ ｉｎ Ｋｙｏｔｏ ｕｎｉｖｅｒｓｉｔｙ.」であり、解は、「京大　名詞」である。

【００４７】解析に用いる情報である素性として、次のものを用いる。１．前の単語自体、２．前の単語の品詞、３．後の単語自体、４．後の単語の品詞、５．解析する単語自体、６．解析する単語のとりうる品詞、７．解の単語、８.解の品詞、９．日本語文と英語文の単語の一致数、１０．英語単語列、１１．解析する単語のとりうる品詞に解の品詞が含まれるか。

【００４８】前記解データのデータベースから抽出される素性は、情報が無い時を＜ｎｏｎｅ＞で記述すると、１．＜ｎｏｎｅ＞、２．＜ｎｏｎｅ＞、３．で（次の単語が「で」だけであることは既存の形態素解析システムで特定してもよいし、もとの対訳データにその情報があることにしてもよい）、４．助詞（前記と同じ）、５．きょうだい、６．名詞（単語辞書を調べてとりうる品詞を探す。複数の品詞をとりうる場合もある）、７．京大、８．名詞、９．３（「待つ－ｗａｉｔ」「で－ｉｎ」「京大－Ｋｙｏｔｏ ｕｎｉｖｅｒｓｉｔｙ」の３つが一致する。この単語の一致の算出は現在解析対象としている部分（きょうだい）も含めて行う。また、現在解析対象としている部分については、解に記述している単語（京大）を用いて行う。日英で単語が一致するかどうかは対訳辞書を用いて行う）、１０．「Ｉ」「ｗａｉｔ」「ｉｎ」「Ｋｙｏｔｏ」「ｕｎｉｖｅｒｓｉｔｙ」（各単語が素性となる）、１１．含まれる。

【００４９】前記では解を用いているので、解と素性の集合の組は正例となる。ここで、「きょうだい」を「兄弟　名詞」とする解候補を用いた解候補と素性の集合の組である負例を考える。この場合、素性は次のようになる。１．＜ｎｏｎｅ＞、２．＜ｎｏｎｅ＞、３．で（次の単語が「で」だけであることは既存の形態素解析システムで特定してもよいし、もとの対訳データにその情報があることにしてもよい）、４．助詞（前記と同じ）、５．きょうだい、６．名詞（単語辞書を調べてとりうる品詞を探す。複数の品詞をとりうる場合もある）、７．兄弟、８．名詞、９．２（「待つ－ｗａｉｔ」、「で－ｉｎ」の２つが一致する。この単語の一致の算出は現在解析対象としている部分（きょうだい）も含めて行う。また、負例の場合は、現在解析対象としている部分については、解に記述していない他の候補の単語（兄弟）を用いて行う）、１０．「Ｉ」「ｗａｉｔ」「ｉｎ」「Ｋｙｏｔｏ」「ｕｎｉｖｅｒｓｉｔｙ」（各単語が素性となる）、１１．含まれる。

【００５０】前記情報を用いて、どのような場合に正例、負例になりやすいかを求めると、「日本語文と英語文の単語の一致数が多い」場合に正例になりやすく、「Ｋｙｏｔｏという素性と京大という素性が共起する」場合に正例になりやすいということを学習する。

【００５１】このような英語の情報をうまく利用したものから、「次の単語が助詞の『で』で」、解の品詞が『名詞』である」場合に正例になりやすいという日本語もともとの性質を利用した学習も同時に行える。

【００５２】「Ｋｙｏｔｏという素性と京大という素性が共起する」場合に正例になりやすいといったことは、素性の共起を自動で考慮することができる機械学習システム、例えば、サポートベクトルマシン、を利用することで可能になる。また、素性の共起を自動で考慮しない学習アルゴリズムの場合は、人手で予めそのような共起を意味する素性を設定しておけばよい。

【００５３】次に、入力する対訳データとして、「＜とうだい＞にいく。」と「Ｉ ｇｏｔｏ Ｔｏｋｙｏ ｕｎｉｖｅｒｓｉｔｙ.」が与えられ、そのうち「とうだい」の部分を形態素解析するように指示されたとする。この時、単語辞書等を調べて解の候補として、「灯台　名詞」と「東大　名詞」があがったとする。

【００５４】先ず、「灯台　名詞」を対象とする。素性は、前記と同様な記述の仕方で、１．＜ｎｏｎｅ＞、２．＜ｎｏｎｅ＞、３．に（次の単語が「に」だけであることは既存の形態素解析システムで特定してもよいし、もとの入力として与えられる対訳データにその情報があることにしてもよい）、４．助詞（前記同様）、５．とうだい、６．名詞（単語辞書を調べてとりうる品詞を探す。複数の品詞をとりうる場合もある）、７．灯

9

台、８．名詞、９．２（「行く－ｇｏ」、「に－ｔｏ」の２つが一致する。この単語の一致の算出は現在解析対象としている部分（とうだい）も含めて行う。また、現在解析対象としている部分については、解の候補の単語（灯台）を用いて行う）、１０．「Ｉ」「ｇｏ」「ｔｏ」「ｕｎｉｖｅｒｓｉｔｙ」（各単語が素性となる）、１１．含まれる。

【００５５】次に、「東大　名詞」を対象とする。素性は、１．＜ｎｏｎｅ＞、２．＜ｎｏｎｅ＞、３．に（次の単語が「に」だけであることは既存の形態素解析システムで特定してもよいし、もとの入力として与えられる対訳データにその情報があることにしてもよい）、４．助詞（前記同様）、５．とうだい、６．名詞（単語辞書を調べてとりうる品詞を探す。複数の品詞をとりうる場合もある）、７．東大、８．名詞、９．２（「行く－ｇｏ」「に－ｔｏ」「東大－Ｔｏｋｙｏ　ｕｎｉｖｅｒｓｉｔｙ」の３つが一致する。この単語の一致の算出は現在解析対象としている部分（とうだい）も含めて行う。また、現在解析対象としている部分については、解の候補の単語（東大）を用いて行う）、１０．「Ｉ」「ｇｏ」「ｔｏ」「ｕｎｉｖｅｒｓｉｔｙ」（各単語が素性となる）、１１．含まれる。

【００５６】ここで学習した結果と前記素性を用いて「灯台　名詞」と「東大　名詞」の正例である確率を求める。「日本語文と英語文の単語の一致数が多い」場合に正例になりやすいといった学習結果により、「東大　名詞」が正例である確率は、「灯台　名詞」のものよりも大きくなり、確率が大きい方の「東大　名詞」が解として出力される。ここでは、一単語を解析の対象とする場合のものを示した。

【００５７】一文全体を解析の対象とする場合は、各単語ごとの解析を組み合わせることで実現できる。この場合、対訳データである「＜とうだい＞にいく。」と「Ｉ　ｇｏ　ｔｏ　Ｔｏｋｙｏ　ｕｎｉｖｅｒｓｉｔｙ．」の解は、「東大　名詞、に助詞、行く　動詞」である。この時、解の候補を複数作り、その中から予め設定した評価値のよいものを選べばよい。解の候補は、「東大　名詞、に　助詞、行く　動詞」「灯台　名詞、に　助詞、行く　動詞」「と　接続詞、宇内　名詞、に　助詞、行く　動詞」である。予め設定する評価値は、各単語ごとに前記方法で前記正例である確率を求め、それの一文全体での積とするとよい。この方法を高速に実現する手段として、ビタビアルゴリズム、ビームサーチが知られている。

【００５８】また、前方から解析する方法をとる場合、前記素性の３に（次の単語が「に」だけであることは既存の形態素解析システムで特定してもよいし、もとの入力として与えられる対訳データにその情報があることにしてもよい）などの表現のうち、「もとの入力として与えられる対訳データにその情報があることにしてもよ

10

い」としていたが、前方から解析する方法をとる場合は、前方の解析結果があるため、「入力として与えられる対訳データにその情報」があることになる。また、後方のものも未だ解析していなったとしても、候補を複数あげ、その一つ一つを解と仮定して解いていく場合は、後方の情報も「入力として与えられる対訳データにその情報」があることになる。

【００５９】尚、解を「単語　品詞」としてそれの組み合わせによって一文全体の解を得る方法の他に、一文全体の解そのものを解として扱ってもよい。

【００６０】（実施例２）英語の構文解析の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【００６１】英語と日本語の２つの対訳データの場合を考える。前記解データは、英語と日本語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。対訳データは、「Ｓｈｅ　ｍｅｔ　ａ　ｂｏｙ　＜ｗｉｔｈ　ａ　ｐｉｃｔｕｒｅ＞．」と「彼女は絵を持っている少年とあった。」であり、解は、「係り先　『ａ　ｂｏｙ』」である。尚、対訳データ中で対象となるフレーズは“＜”、“＞”の記号で囲んでおく。

【００６２】解析に用いる素性は、１．解析するフレーズの意味的主辞の単語、２．解析するフレーズの意味的主辞の単語の意味カテゴリ、３．解析するフレーズの構文的主辞の単語、４．解析するフレーズの構文的主辞の単語の意味カテゴリ、５．解の係り先の単語、６．解の係り先の単語の意味カテゴリ、７．解の係り先の単語の品詞、８．日本語単語列、９．日本語文と英語文の対応する２つのフレーズのかかり受けの一致数、である。

【００６３】前記から抽出される素性は、１．ｐｉｃｔｕｒｅ（「ｗｉｔｈ　ａ　ｐｉｃｔｕｒｅ」の意味主辞は名詞句の主辞の「ｐｉｃｔｕｒｅ」となる。フレーズのどの部分が意味主辞で、どの部分が構文的主辞になるかは、予め文法を用いて定めておくとよい）、２．製品（どういう単語がどういう意味カテゴリになるかは、単語意味辞書を用いることで特定できる）、３．ｗｉｔｈ、４．前置詞、５．ｂｏｙ、６．人、７．名詞、８．「彼女」「は」「絵」「を」「もっている」「少年」「と」「あった」（各単語が素性となる。また、これは既存の形態素解析システムで分割してもよいし、もとの対訳データで分割されたものが与えられていたとしてもよいし、もとの対訳データで分割されたものが与えられていたとしてもよい）、９．４（「ｓｈｅ－ｍｅｔ」と「彼女は－あった」、「ｍｅｔ－ａｂｏｙ」と「少年と－あった」、「ｗｉｔｈ－ａ　ｐｉｃｔｕｒｅ」と「絵を－もっている」、「ａ　ｂｏｙ－ｗｉｔｈ」と「もっ

ている－少年」の４つ。解析対象の「ｗｉｔｈ」の係り先「ａ　ｂｏｙ」との関係以外のフレーズの係り受けの情報は既存の構文解析システムで特定してもよいし、もとの対訳データにその情報があることにしてもよい）。これらは、解の部分を用いているので、正例となる。

【００６４】次に、負例として係り先を「ｍｅｔ」としたものを考える。この場合、素性は、１．ｐｉｃｔｕｒｅ（「ｗｉｔｈ　ａ　ｐｉｃｔｕｒｅ」の意味主辞は名詞句の主辞の「ｐｉｃｔｕｒｅ」となる。フレーズのどの部分が意味主辞で、どの部分が構文的主辞になるかは、予め文法を用いて定めておくとよい）、２．製品（どういう単語がどういう意味カテゴリになるかは、単語意味辞書を用いることで特定できる）、３．ｗｉｔｈ、４．前置詞、５．ｍｅｅｔ、６．知覚動詞、７．動詞、８．「彼女」「は」「絵」「を」「もっている」「少年」「と」「あった」（各単語が素性となる。また、これは既存の形態素解析システムで分割してもよいし、もとの対訳データで分割されたものが与えられていたとしてもよいし、もとの対訳データで分割されたものが与えられていたとしてもよい）、９．３（「ｓｈｅ－ｍｅｔ」と「彼女は－あった」、「ｍｅｔ－ａ　ｂｏｙ」と「少年と－あった」、「ｗｉｔｈ－ａ　ｐｉｃｔｕｒｅ」と「絵を－もっている」の３つ。解析対象のｗｉｔｈの係り先ａ　「ｂｏｙ」との関係以外のフレーズの係り受けの情報は、既存の構文解析システムで特定してもよいし、もとの対訳データにその情報があることにしてもよい）、となる。

【００６５】前記情報を用いて、どういう場合に正例あるいは負例になりやすいかを求めると、「日本語文と英語文の構文リンクの一致数が多い」場合に正例になりやすいとか、「『もっている』という日本語単語素性と解析するフレーズの構文的主辞の単語素性『ｗｉｔｈ』と、解の係り先の単語の品詞素性『名詞』が共起する」場合に正例になりやすいとかを学習できる。このような日本語の情報をうまく利用するものの他に、英語の素性を用いるため、英語もともとの性質を利用した学習も同時に行える。

【００６６】例えば、「製品を意味主辞にもつ『ｗｉｔｈ』のフレーズは、『ｍｅｅｔ』よりも名詞にかかりやすい」など。尚、「『もっている』という日本語単語素性と解析するフレーズの構文的主辞の単語素性『ｗｉｔｈ』と、解の係り先の単語の品詞素性『名詞』が共起する」は、「ｗｉｔｈ」を「もっている」と和訳する場合は、そのフレーズは名詞にかかりやすいということを意味する。

【００６７】次に、入力する英語と日本語の対訳データとして、「Ｓｈｅ　ｌｏｏｋｅｄａｔ　ａ　ｂｏｙ　＜ｗｉｔｈ　ａ　ｔｅｌｅｓｃｏｐｅ＞．」「彼女は望遠鏡を持っている少年を見た。」が与えられ、そのうち「ｗｉｔｈ　ａ　ｔｅｌｅｓｃｏｐｅ」の部分の係り先を求めるように指示されたとする。尚、「Ｓｈｅ　ｌｏｏｋｅｄ　ａｔ　ａ　ｂｏｙ　ｗｉｔｈ　ａ　ｔｅｌｅｓｃｏｐｅ．」は構文的に曖昧な表現であり、「ｗｉｔｈ　ａ　ｔｅｌｅｓｃｏｐｅ」は「ｌｏｏｋｅｄ」にも「ａｔ　ａ　ｂｏｙ」にもかかりうる。「ｌｏｏｋｅｄ」にかかる場合は、「彼女は望遠鏡で少年を見た。」の意味になる。

【００６８】ここまでのフレーズのまとめあげの結果などから、解の候補として「ｌｏｏｋｅｄ」と「ａ　ｂｏｙ」があがったとして、先ず「ｌｏｏｋｅｄ」を対象とする。素性は、１．ｔｅｌｅｓｏｐｅ（「ｗｉｔｈ　ａ　ｔｅｌｅｓｃｏｐｅ」の意味主辞は名詞句の主辞の「ｔｅｌｅｓｃｏｐｅ」となる。フレーズのどの部分が意味主辞で、どの部分が構文的主辞になるかは予め文法を用いて定めておくとよい）、２．製品（どういう単語がどういう意味カテゴリになるかは単語意味辞書を用いることで特定できる）、３．ｗｉｔｈ、４．前置詞、５．ｌｏｏｋ、６．知覚動詞、７．動詞、８．「彼女」「は」「望遠鏡」「を」「もっている」「少年」「見た」（各単語が素性となる。またこれは既存の形態素解析システムで分割してもよいし、もとの対訳データで分割されたものが与えられていたとしてもよい）、９．３（「ｓｈｅ－ｌｏｏｋｅｄ」と「彼女は－見た」、「ｌｏｏｋｅｄ－ａｔ　ａ　ｂｏｙ」と「少年を－見た」、「ｗｉｔｈ－ａ　ｔｅｌｅｓｃｏｐｅ」と「望遠鏡を－もった」の３つ。解析対象の「ｗｉｔｈ」の係り先「ａｂｏｙ」との関係以外のフレーズの係り受けの情報は既存の構文解析システムで特定してもよいし、もとの対訳データにその情報があることにしてもよい）、である。

【００６９】次に、「ａ　ｂｏｙ」を対象とする。素性は、１．ｔｅｌｅｓｏｐｅ（「ｗｉｔｈ　ａ　ｔｅｌｅｓｃｏｐｅ」の意味主辞は名詞句の主辞の「ｔｅｌｅｓｃｏｐｅ」となる。フレーズのどの部分が意味主辞で、どの部分が構文的主辞になるかは予め文法を用いて定めておくとよい）、２．製品（どういう単語がどういう意味カテゴリになるかは、単語意味辞書を用いることで特定できる）、３．ｗｉｔｈ、４．前置詞、５．ｂｏｙ、６．人、７．名詞、８．「彼女」「は」「望遠鏡」「を」「もっている」「少年」「見た」（各単語が素性となる。またこれは既存の形態素解析システムで分割してもよいし、もとの対訳データで分割されたものが与えられていたとしてもよい）、９．４（「ｓｈｅ－ｌｏｏｋｅｄ」と「彼女は－見た」、「ｌｏｏｋｅｄ－ａｔ　ａ　ｂｏｙ」と「少年を－見た」、「ｗｉｔｈ－ａ　ｔｅｌｅｓｃｏｐｅ」と「望遠鏡を－もった」、「ａ　ｂｏｙ－ｗｉｔｈ」と「もった－少年」の４つ。解析対象の「ｗｉｔｈ」の係り先「ａｂｏｙ」との関係以外のフレーズの係り受けの情報は既存の構文解析システムで特定してもよいし、もとの対訳データにその情報があることにしてもよい）、である。

13

【００７０】ここで学習した結果と前記素性を用いて「ｌｏｏｋｅｄ」と「ａ　ｂｏｙ」の正例である確率を求める。「日本語文と英語文の構文リンクの一致数が多い」場合に正例になりやすいとか、「『もっている』という日本語単語素性と解析するフレーズの構文的主辞の単語素性『ｗｉｔｈ』と解の係り先の単語の品詞素性『名詞』が共起する」場合に正例になりやすいといった学習結果により、「ａ　ｂｏｙ」が正例である確率は、「ｏｏｋｅｄ」のものよりも大きくなり、確率が大きい方の「ａ　ｂｏｙ」が解として出力される。

【００７１】ここでは、例えば、構文リンクの一致数を求める前記素性９が思うように動かなかったとしても（素性９は既存の構文解析システムなどを前提にするため、場合によっては動かない可能性がある）、「『もっている』という日本語単語素性と解析するフレーズの構文的主辞の単語素性「ｗｉｔｈ」と解の係り先の単語の品詞素性『名詞』が共起する」場合に正例になりやすいという方の性質の方をうまく使うことで、「ａ　ｂｏｙ」を正しく解として出力する。

【００７２】機械学習手法は情報が一部不足した場合も他の情報をうまく利用することができる。ここでは、１つのフレーズの係り先の特定をするものを示した。構文解析でも形態解析と同様、一文全体を解析の対象とする場合は、各フレーズでの解析を組み合わせることで実現できる。

【００７３】（実施例３）日本語の格解析の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【００７４】日本語と英語の２つの対訳データの場合を考える。前記解データは、日本語と英語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。対訳データは、「みかん＜も＞食べた。」と「Ｗｅ　ａｔｅ　ｏｒａｎｇｅｓ,ｔｏｏ．」であり、解は、「格　を」である。

【００７５】日本語の格解析とは、「は」「も」などの副助詞で表現されたり、連体節で表現されて（例、「食べたみかん」）、ガ格、ヲ格などの格助詞が消えている場合に、その消えた格を推定することを意味する。また、ここで求める格を意味関係にもつ体言と用言を単に体言と用言と書く。

【００７６】解析に用いる素性は、１．体言の単語自体、２．体言の単語の意味カテゴリ、３．用言の単語自体、４．用言の単語の意味カテゴリ、５．英語の単語２－ｇｒａｍ列（２－ｇｒａｍとは２連続表現を意味し、単語２－ｇｒａｍは単語が２つ連続する表現を意味する）、６．体言－用言に対応する英語表現の構文パターン、である。

14

【００７７】前記から抽出される素性は、１．みかん、２．食べ物、３．食べる、４．飲食関係の動詞、５．「Ｗｅ　ａｔｅ」「ａｔｅ　ｏｒａｎｇｅｓ」「ｏｒａｎｇｅｓ,ｔｏｏ」、６．ＶＰ　ＮＰ（英語文を既存の構文解析システムなどでフレーズパターンを出力できるようにする。また、日本語に対する表現の特定は、日英翻訳辞書で単語逐語訳をして行う。もしくは、前記情報をまとめて、もとの対訳データで与えられるとしてもよい）、である。

【００７８】前記素性の情報と解の情報「を」を利用して、機械学習手法により、どういう場合に「を」になりやすく、どういう場合に「が」になりやすいかなどを学習する。具体的には、英語表現の構文パターンが「ＶＰ　ＮＰ」のときに「を」になりやすいとか、「ＮＰ　ＶＰ」のときに「が」になりやすい、などを学習する。また、もとの日本語だけの情報に基づく体言「食べ物」、用言「飲食関係の動詞」のときに「を」になりやすいということも同時に学習する。

【００７９】次に、入力する対訳データとして、「本＜は＞読んだ。」と「Ｗｅ　ｒｅａｄｔｈｅ　ｂｏｏｋ．」が与えられ、そのうち「は」の部分を格解析するように指示されたとする。この時、素性は、１．本、２．製品、３．読む、４．文書関係の動詞、５．「Ｗｅ　ｒｅａｄ」「ｒｅａｄ　ｔｈｅ」「ｔｈｅ　ｂｏｏｋ」、６．ＶＰ　ＮＰ、である。

【００８０】前記素性でどの格になりやすいかを推定する。「ＶＰ　ＮＰ」の素性の存在で「を」になりやすいと判定し、それが解として出力される。ところで、「ＶＰＮＰ」ならばいつでも「を」とは限らない。例えば、「Ｉ　ｌｉｋｅ　ａｐｐｌｅｓ．」だと、「りんごが好き」で「が」である。このような例外的現象も機械学習手法であると簡単に学習できる。例えば、前記素性３の用言が「好き」の場合は、「ＶＰ　ＮＰ」でも、「が」と判定するように学習することになる。

【００８１】（実施例４）単文の時制及びモダリティ表現の推定の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【００８２】日本語と英語の２つの対訳データの場合を考える。前記解データは、日本語と英語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。対訳データは、「京大で待つ。」と「Ｉ　ｗａｉｔ　ｉｎ　Ｋｙｏｔｏ　ｕｎｉｖｅｒｓｉｔｙ．」であり、解は、「現在」であり、また対訳データは、「京大に行く。」と「Ｉ　ｇｏ　ｔｏ　Ｋｙｏｔｏ　ｕｎｉｖｅｒｓｉｔｙ．」であり、解は、「未来」である。ここでは、現在と未来しかあげていないが、過去、完了、要望、可能など、種種の

15

分類が考えられる。これらの分類は、文法書などを参考に予め定めておく。

【００８３】前記と同様にして、素性は、１．日本語文末文字列、２．英語主節の動詞句表現の単語列、３．日本語単語列、４．英語単語列、である。

【００８４】対訳データが、「京大に行く。」と「Ｉ ｗｉｌｌ ｇｏ ｔｏ Ｋｙｏｔｏｕｎｉｖｅｒｓｉｔｙ．」であり、解が、「未来」である場合、抽出される素性は、１．「く」「行く」など、２．「ｗｉｌｌ ｇｏ」「ｇｏ」「ｗｉｌｌ」、３．「京大」「で」「待つ」、４．「Ｉ」「ｗｉｌｌ」「ｇｏ」「ｔｏ」「Ｋｙｏｔｏ」「ｕｎｉｖｅｒｓｉｔｙ」である。

【００８５】日本語だけでは「未来」などの時制を特定するのは難しいが、英語の前記素性２「ｗｉｌｌ」などがあると「未来」などの時制を特定するのは容易である。また、英語主節の動詞句表現の特定はなんらかの構文解析システムが必要になる。場合によっては、その解析結果が間違う可能性もある。そのような場合は、日本語の文末表現の情報も使うことで、場合によっては、英語側の情報が誤っても日本語の方の情報でうまく行く場合がある。尚、実際の解析は省略する。

【００８６】（実施例５）名詞句の指示性の推定の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。名詞句指示性には、総称名詞句、定名詞句、不定名詞句があり、またこれを特定することで冠詞の生成などに役立つ。

【００８７】日本語と英語の２つの対訳データの場合を考える。前記解データは、日本語と英語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。対訳データは、「<犬>がいる。」と「Ｔｈｅｒｅ ｉｓ ＜ａ ｄｏｇ＞．」であり、解は、「不定名詞」であり、また対訳データは、「<その犬>は役に立つ。」と「Ｔｈｅ ｄｏｇ ｉｓ ｕｓｅｆｕｌ」であり、解は、「定名詞」であり、また対役データは、「<犬>は役に立つ。」と「Ｔｈｅ ｄｏｇ ｉｓ ｕｓｅｆｕｌ．」であり、解は、「総称名詞」である。

【００８８】前記と同様にして、素性は、１．日本語周辺表層表現、２．英語周辺表層表現、である。

【００８９】対訳データが、「<その犬>は役に立つ。」と「Ｔｈｅ ｄｏｇ ｉｓ ｕｓｅｆｕｌ．」であり、解が、「総称名詞」である場合、抽出される素性は、１．「その」「役に立つ」など、２．「Ｔｈｅ」「ｉｓ」「ｕｓｅｆｕｌ」、である。

【００９０】英語があると、冠詞が、定冠詞か否か、不定冠詞かで、「定名詞」「不定名詞」のどちらの可能性もないことが分かる。また、日本語で「その犬」のように「その」が存在していると「総称名詞」の可能性はなくなる。そのような学習は、予め素性を適切に定めておくと機械学習で自動で行うことができる。尚、実際の解析は省略する。

16

【００９１】（実施例６）単文の時制及びモダリティ表現の中国語への翻訳の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【００９２】日本語と英語の２つの対訳データの場合を考える。前記解データは、日本語と英語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。

【００９３】対訳データは、「あなたたちはあの映画を見ましたか？」と「Ｈａｖｅ ｙｏｕ ｓｅｅｎ ｔｈａｔ ｆｉｌｍ?」であり、解は、「看了」（「看」が見る、「了」が「～した」を意味する）であり、また対訳データは、「ここで写真をとってもよいですか？」と「Ｍａｙ Ｉ ｈａｖｅ ａ ｐｉｃｔｕｒｅ ｈｅｒｅ？」であり、解は、「可以照」（「照」が「写真をとる」、「可以」が「～してよい」を意味する）である。

【００９４】これは、実施例４の分類カテゴリを単純に中国語の動詞表現にしただけである。実施例４と同様に機械学習を用いると、日本語と英語の情報をうまく組み合わせて用いることができる。

【００９５】また、ここでは動詞句表現を分類としたが、動詞句表現で使われる助動詞だけをとりあえず推定し、それを後の処理と組み合わせて翻訳に利用することも可能である。例えば、対訳データが、「あなたたちはあの映画をみましたか？」と「Ｈａｖｅ ｙｏｕ ｓｅｅｎ ｔｈａｔ ｆｉｌｍ?」であり、解が、「了」であり、対訳データが、「ここで写真をとってもよいですか?」と「Ｍａｙ Ｉｈａｖｅ ａ ｐｉｃｔｕｒｅ ｈｅｒｅ？」であり、解が、「可以」であるなど。尚、実際の解析は省略する。

【００９６】（実施例７）英語単語の中国語単語への翻訳の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【００９７】英語と日本語の２つの対訳データの場合を考える。前記解データは、英語と日本語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。

【００９８】対訳データは、「Ｍａｙ Ｉ ｈａｖｅ ＜ａ ｐｉｃｔｕｒｅ＞ ｈｅｒｅ?」と「ここで写真をとってもよいですか？」であり、解は、「相」（「相」は「写真」を意味する）である。

17

【００９９】英単語「ａ　ｐｉｃｔｕｒｅ」の意味には写真の他にも絵の意味があり、絵を意味する中国語単語「画儿」との訳し訳の必要がある。日本語の「写真」という語の存在のおかげで正しく「相」と翻訳できる。動詞句が単語になっただけで実施例６とほぼ同様に扱える。尚、実際の解析は省略する。

【０１００】（実施例８）中国語への翻訳における生成される中国語での構文構造推定の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【０１０１】日本語、英語及び中国語の３つの対訳データの場合を考える。前記解データは、日本語、英語及び中国語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。

【０１０２】対訳データは、「私は炒飯を食べます。」、「Ｉ　ｈａｖｅ　ｆｒｉｅｄ　ｒｉｃｅ．」及び（中国語単語逐語訳列）「我＜炒飯＞吃」であり、解は、「炒飯」の係り先「吃」である。対訳データ中で係り先を求める中国語単語は“＜”“＞”の記号で囲んでおく。

【０１０３】解析に用いる情報である素性は、１．解析するフレーズの主辞の単語、２．解析するフレーズの主辞の単語の意味カテゴリ、３．解析するフレーズの主辞の単語の品詞、４．解の係り先の単語、５．解の係り先の単語の意味カテゴリ、６．解の係り先の単語の品詞、７．日本語単語列、８．英語単語列、９．日本語文での係り受けと対応するか、１０．英語文での係り受けと対応するか、である。

【０１０４】前記解データのデータベースから抽出される「吃」を解とする場合の素性は、前記同様に表すと、１．炒飯、２．食べ物、３．名詞、４．吃、５．飲食関係の動詞、６．動詞、７．「私」「は」「炒飯」「を」「食べます」、８．「Ｉ」「ｈａｖｅ」「ｆｒｉｅｄ　ｒｉｃｅ」、９．対応する（日本語では「炒飯を」「食べます」にかかっている。また、この種の情報は既存のシステムで求めてもようし、入力で与えられているとしてもよい）、１０．対応する（英語では「ｆｒｉｅｄ　ｒｉｃｅ」が「ｈａｖｅ」にかかっている）、である。これは正例となる。

【０１０５】また、「我」を解とする場合の素性は、１．炒飯、２．食べ物、３．名詞、４．我、５．人、６．名詞、７．「私」「は」「炒飯」「を」「食べます」、８．「Ｉ」「ｈａｖｅ」「ｆｒｉｅｄ　ｒｉｃｅ」、９．対応しない、１０．対応しない、である。これは負例となる。

【０１０６】以上の情報で学習すると、日本語文若しくは英語文での係り受けと対応すると正例である確率が高くなるように学習することになる。また、構文構造が対

18

応とれないようにしか解析できない場合もある。また、日本語、英語の構文構造の解析を失敗する場合もある。そのような場合は、前記１～８などの他の素性が役に立つことになる。尚、実際の解析は省略する。

【０１０７】（実施例９）中国語への一文翻訳の場合、図１に示す解－素性対抽出部１２において、解と素性の集合及び解候補と素性の集合の組を抽出する方法を利用して解く。

【０１０８】日本語と英語の２つの対訳データの場合を考える。前記解データは、日本語と英語の対訳データに解の情報が付与されたものであるので、以下の対訳データの場合に、解データのデータベースの構成は、次のようになる。尚、対訳データ中で対象となる単語を“＜”、“＞”の記号で囲っておく。

【０１０９】対訳データは、「私は炒飯を食べます。」と「Ｉ　ｈａｖｅ　ｆｒｉｅｄ　ｒｉｃｅ．」であり、解は、「我吃炒飯」である。解の部分には翻訳結果がはいる。

【０１１０】前記素性の組を抽出する方法では、解の候補を作成する必要がある。簡単な方法としては、あらゆる単語逐語訳と、またそれの語順をあらゆる場合で並べ替えたものを全て解の候補とすればよい。そしてその中から、正例の確率が最も大きいものを選ぶとよい。

【０１１１】また、この方法で、解の候補の数が発散する場合には、問題を部分部分に分割し、各部分で正例の確率を算出し、その積が最大になるように部分部分を統合するようにしておけばよい。これは、一文全体の形態素解析を行うのと同様である。

【０１１２】（実施例１０）機械学習手法の場合、教師信号が同じ形をしているものは併用して学習できる。例えば、実施例１の形態素解析では、対訳データは、「＜きょうだい＞で待つ。」と「Ｉ　ｗａｉｔ　ｉｎ　ｋｙｏｔｏ　ｕｎｉｖｅｒｓｉｔｙ．」であり、解は、「京大　名詞」といった形のデータを教師信号として用いるが、「＜きょうだい＞で待つ。」「京大　名詞」のような対訳データでないものも教師信号と扱える。この場合、英語に関係する素性情報に欠けるが、その部分は無かったとしても日本語に関係する素性情報が残るので、その情報を使って学習することになる。

【０１１３】ところで、対訳データが、「＜きょうだい＞で待つ。」と「Ｉ　ｗａｉｔ　ｉｎ　ｋｙｏｔｏ　ｕｎｉｖｅｒｓｉｔｙ．」であり、解が、「京大　名詞」であるものに、更に形態素情報もふったコーパスはあまりみないが、「＜きょうだい＞で待つ。」と「京大　名詞」のような形態素情報もふったコーパスは多く存在する。このデータも使って学習できると、学習データが多いため精度が向上する。

【０１１４】また、前記併用型の場合、大規模に使える単言語の学習データと、情報量の多い２言語対訳の学習データを併用するので、非常に強力である。また、既存

19

のシステムで学習を用いるものは、単言語の学習データを用いているので、この併用型は少なくとも既存の学習システムと同等程度の能力を確保した上で、さらに２言語対訳の学習データを併用するというものになっている。

【０１１５】

【発明の効果】本発明によれば、大量の変換規則を用意する必要がなく、複数の自然言語で記述された処理対象文から他の自然言語及び／又は同じ自然言語で記述された処理結果文への言語変換及び／又は複数の言語における言語解析を行うことができる複数言語入力での言語処理方法及び言語処理装置が得られる。

【図面の簡単な説明】

【図１】本発明による複数言語入力での言語処理装置の実施形態を示すブロック図である。

【図２】本発明による複数言語入力での言語処理方法の実施形態を示すフローチャートである。

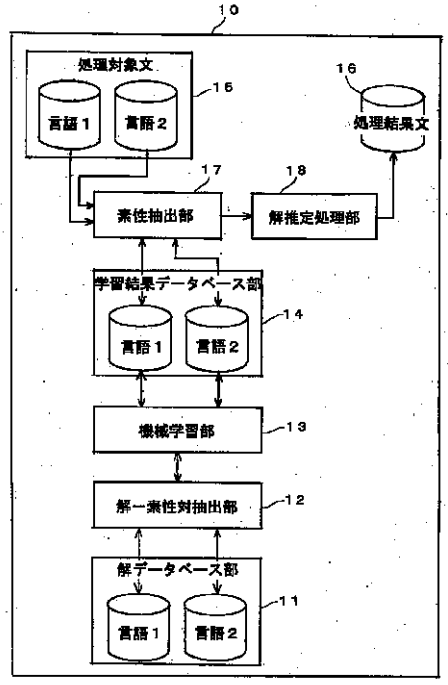
【図３】従来の機械学習手法を用いた単言語入力での言語処理装置のブロック図である。

20

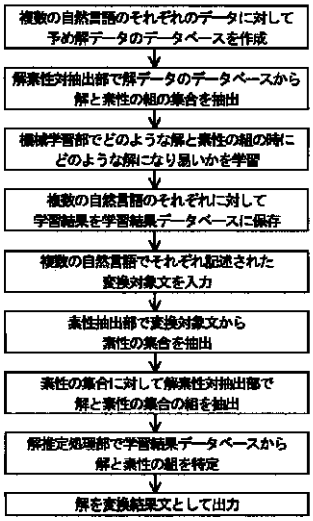
【符号の説明】

１０　言語処理装置
１１　解データベース部
１２　解－素性対抽出部
１３　機械学習部
１４　学習結果データベース部
１５　処理対象文
１６　処理結果文
１７　素性抽出部
１８　解推定処理部
３０　言語処理装置
３１　解データベース部
３２　解－素性対抽出部
３３　機械学習部
３４　学習結果データベース部
３５　変換対象文
３６　素性抽出部
３７　解推定処理部
３８　変換結果文

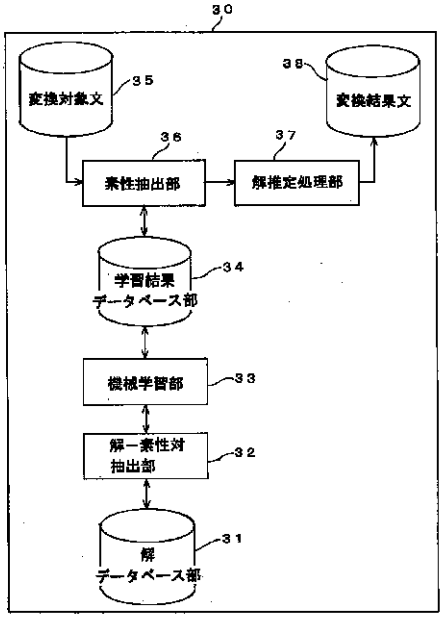
【図１】



【図２】



【図３】


30
変換対象文
35
38
変換結果文
36
37
素性抽出部
解推定処理部
学習結果
データベース部
34
機械学習部
33
解－素性対
抽出部
32
解
データベース部
31