(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第5146979号
(P5146979)
(45)発行日　平成25年2月20日(2013.2.20)　(24)登録日　平成24年12月7日(2012.12.7)

(51)Int.Cl.　F I
G06F 17/27 (2006.01)　G06F 17/27 M

請求項の数 6　(全 21 頁)

(21)出願番号　特願2006-154497(P2006-154497)
(22)出願日　平成18年6月2日(2006.6.2)
(65)公開番号　特開2007-323475(P2007-323475A)
(43)公開日　平成19年12月13日(2007.12.13)
審査請求日　平成20年12月8日(2008.12.8)

(73)特許権者　393031586
株式会社国際電気通信基礎技術研究所
京都府相楽郡精華町光台二丁目２番地２
(73)特許権者　000208891
ＫＤＤＩ株式会社
東京都新宿区西新宿二丁目３番２号
(74)代理人　100099933
弁理士　清水　敏
(72)発明者　隅田　英一郎
京都府相楽郡精華町光台二丁目２番地２
株式会社国際電気通信基礎技術研究所内
(72)発明者　菅谷　史昭
埼玉県ふじみ野市大原二丁目１番１５号
株式会社ＫＤＤＩ研究所内

最終頁に続く

(54)【発明の名称】自然言語における多義解消装置及びコンピュータプログラム

(57)【特許請求の範囲】
【請求項１】
自然言語文からなる入力文において、ある単語と、前記入力文において前記ある単語が置かれた文脈と、前記ある単語の意味を表す可能性のある複数の意味候補を含む意味候補の集合とが与えられると、当該意味候補の集合の中から、前記文脈において前記ある単語の意味として最も適切なものを選択する、自然言語における多義解消装置であって、
前記ある単語と、前記意味候補集合中の意味候補との組合せの各々について、所定のコーパスから、当該組合せを構成する語同士が共起する文書の集合を収集するための文書収集手段と、
前記文書収集手段によって前記組合せの各々について収集された前記文書の集合を学習データとして用い、前記ある単語と、その単語の文書中の文脈とが与えられると、前記ある単語の当該文脈中での意味として最適な意味候補を前記意味候補集合中から選択する分類器を自動的に作成するための分類器作成手段と、
前記入力文において、前記ある単語が置かれた文脈に基づいて、前記ある単語の意味として最適なものを、前記意味候補の集合の中から前記分類器を用いて選択するための分類実行手段とを含み、
前記文書収集手段は、前記ある単語と、前記意味候補集合中の意味候補との組合せの各々について、前記ある単語とその意味候補との両方を検索キーワードとして、インターネット上に存在する文書からなる仮想的コーパスから、当該組合せを構成する語同士が共起する文書の集合を検索し収集するための検索手段を含み、

前記分類器作成手段は、

前記文書収集手段によって前記組合せの各々について収集された前記文書の集合のうち、集合に含まれる文書の数が多いものを所定の基準に従って選択し、それら文書の集合に対応する意味候補のみを前記意味候補の集合の要素として選択する処理を行なうための意味候補選択手段と、

前記意味候補選択手段により選択された文書集合を学習データとして用い、前記ある単語と、その単語の文書中の文脈とが与えられると、前記ある単語の当該文脈中での意味として最適な意味候補を、前記意味候補集合中から選択する分類器を機械学習により自動的に作成するための機械学習手段とを含む、自然言語における多義解消装置。

【請求項２】

前記機械学習手段は、

前記文書集合選択手段により選択された文書集合に含まれる文書の各々に対し、当該文書中における前記ある単語の位置の前後の所定範囲に存在する単語列から、当該文書中における前記ある単語の文脈の特徴量を表す、所定の構成の学習用の特徴量ベクトルを算出するための特徴量ベクトル算出手段と、

前記文書集合選択手段により選択された文書集合に含まれる文書の各々に対して前記特徴量ベクトル算出手段により算出された学習用の特徴量ベクトルと、当該文書の検索時に使用された意味候補とを組にして学習用データを作成し、当該学習用データを用いた機械学習により、前記学習用の特徴量ベクトルと同じ構成の分類用の特徴量ベクトルが与えられると、当該分類用の特徴量ベクトルに対応する文脈中における前記ある単語の意味として最適なものを、前記意味候補集合中から選択する所定の分類器を自動的に作成するための手段とを含む、請求項１に記載の自然言語における多義解消装置。

【請求項３】

前記意味候補選択手段は、前記文書収集手段によって前記組合せの各々について収集された前記文書の集合のうち、集合に含まれる文書の数が多い所定の個数の集合を選択し、それら文書の集合に対応する意味候補のみを前記意味候補として選択する処理を行なうための手段を含む、請求項１又は請求項２に記載の、自然言語における多義解消装置。

【請求項４】

前記意味候補選択手段は、前記文書収集手段によって前記組合せの各々について収集された前記文書の集合のうち、集合に含まれる文書の数が予め定められるしきい値より大きな集合を選択し、それら文書の集合に対応する意味候補のみを前記意味候補として選択する処理を行なうための手段を含む、請求項１又は請求項２に記載の、自然言語における多義解消装置。

【請求項５】

前記収集するための手段は、前記ある単語と、前記意味候補集合中の意味候補との組合せの各々について、インターネット上に存在するウェブページからなる仮想的コーパスから、当該組合せを構成する語が共起するウェブページの集合を検索し、所定の定数を上限とした要素数の集合として収集するための手段を含む、請求項１に記載の自然言語における多義解消装置。

【請求項６】

コンピュータにより実行されると、当該コンピュータを、請求項１～請求項５のいずれかに記載の自然言語における多義解消装置として機能させる、コンピュータプログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

この発明は自然言語処理に関し、特に、単語の読み（日本語における仮名表記）、アクロニム（頭字語）のフルスペル、及び二つの言語の間での訳語の対応などに見られるあい

まい性を解消するための自然言語処理に関する。
【背景技術】
【０００２】
自然言語には、あいまい性が常に付きまとう。例えば同形異音語という問題がある。同形異音語とは、読みが複数ある単語のことである。例えば英語の「ｂｏｗ」という単語には、「ｂｏｗ」（蝶型リボン）と「ｂｏｗ」（船首）という二つの読み方がある。日本語でもこうした例は多い。例えば「大平」という語は、「オオヒラ」とも、「タイヘイ」とも、「オオダイラ」とも読める。
【０００３】
こうしたあいまい性は、アクロニムにも存在する。例えば「ＡＣＬ」というアクロニムは、「The Association for Computational Linguistics」、「Anterior Cruciate Ligament」、及び「Access Control List」のいずれとも解釈できる。同様のあいまい性は、翻訳の際の訳語の決め方等にも見出すことができる。
【０００４】
人間の場合、こうしたあいまい性については、その語が生起した状況などに応じて適宜適切に判断をしたり、いずれかの手段を用いて候補をさがし、その中で状況に応じて最も適していると思われるものを選択したりすることで解決している。しかし、自然言語処理でそのような処理を実現することは困難である。
【０００５】
このようなあいまい性は、自然言語処理において重大な問題となり得る。例えば、日本語テキストの読上げにおいて、同形異音語に遭遇した場合、適切な発音で読上げを行なうためには、その発音（かな表記）を決定する必要がある。さもないと、不適切な読上げが行なわれてしまう。
【０００６】
こうした問題を解決するための提案が非特許文献１でなされている。非特許文献１では、予め、ある単語Ｗとその対応する意味Ｓｉとを記述した学習データを人手で用意し、その学習データを用いて、ある単語Ｗが与えられたときに意味Ｓｉのうちで適切なものを選択する分類器を作成する。
【非特許文献１】梅村祥之、清水司、「音声合成システムのための同形異音語の読み分け」、豊田中央研究所Ｒ＆Ｄレビュー、２０００年、第３５巻第１号、６７頁～７４頁
【発明の開示】
【発明が解決しようとする課題】
【０００７】
しかし、非特許文献１に開示された方法では、学習データを人手で用意する必要があり、時間と費用とがかさむという問題がある。また、限られた人の手によって学習データが作成されるので、学習データに偏りが生ずる可能性もあるため、信頼性が低いという問題もある。
【０００８】
それ故に本発明の目的は、自然言語に伴うあいまい性を、容易に、かつ信頼性高く解決できる、自然言語における多義解消装置を提供することである。
【課題を解決するための手段】
【０００９】
本発明の第１の局面によれば、自然言語における多義解消装置は、自然言語文からなる入力文において、ある単語と、入力文においてある単語が置かれた文脈と、ある単語の意味を表す可能性のある複数の意味候補を含む意味候補の集合とが与えられると、当該意味候補の集合の中から、文脈においてある単語の意味として最も適切なものを選択する、自然言語における多義解消装置であって、ある単語と、意味候補集合中の意味候補との組合せの各々について、所定のコーパスから、当該組合せを構成する語が共起する文書の集合を収集するための文書収集手段と、文書収集手段によって組合せの各々について収集された文書の集合を学習データとして用い、ある単語と、その単語の文書中の文脈とが与えら

れると、ある単語の当該文脈中での意味として最適な意味候補を意味候補集合中から選択する分類器を自動的に作成するための分類器作成手段と、入力文において、ある単語が置かれた文脈に基づいて、ある単語の意味として最適なものを、意味候補の集合の中から分類器を用いて選択するための分類実行手段とを含む。

【００１０】

入力文中のある単語と、その単語が置かれた文脈と、その単語に意味を表す可能性のある複数の意味候補を含む意味候補の集合が与えられると、その単語と意味候補との組合わせの各々について、文書収集手段が所定のコーパスから当該組合せを構成する単語が共起する文書の集合を収集する。分類器作成手段は、組合せの各々について収集された文書の集合を学習データとして用いて分類器を作成する。この分類器は、ある単語と、その単語の文書中の文脈とが与えられると、ある単語の当該文脈中での意味として最適な意味候補を意味候補集合中から選択する機能を持つ。分類手段は、入力文中の単語と、その単語が置かれた文脈とを、このようにして作成された分類器に与え、その結果に基づいて、入力文中の単語の意味として最適なものを、意味候補の集合の中から選択する。

【００１１】

すなわち、この装置では、ある単語と、その単語の文脈と、その単語に対応する可能性のある複数の意味候補とが与えられると、文脈から適切と思われる意味候補を自動的に選択できる。この作業には人手を介在させる必要はない。従って、容易に適切な意味候補を選択し、入力された単語の多義性を解消できる多義解消装置を提供できる。

【００１２】

好ましくは、分類器作成手段は、文書収集手段によって組合せの各々について収集された文書の集合のうち、集合に含まれる文書の数が多いものを所定の基準に従って選択し、それら文書の集合に対応する意味候補のみを意味候補の集合の要素として選択する処理を行なうための意味候補選択手段と、意味候補選択手段により選択された文書集合を学習データとして用い、ある単語と、その単語の文書中の文脈とが与えられると、ある単語の当該文脈中での意味として最適な意味候補を、意味候補集合中から選択する分類器を機械学習により自動的に作成するための機械学習手段とを含む。

【００１３】

収集された文書集合のうち、集合に含まれる文書の数が少ないものは意味候補選択手段により棄却される。集合に含まれる文書の数が少ないということは、その単語と、その集合に対応する意味候補とが共起する可能性が他と比較して少ないということである。従って、与えられた文脈におけるある単語の意味として不適切なものを排除できる。その結果、分類の信頼性を高めることができる。

【００１４】

さらに好ましくは、機械学習手段は、文書集合選択手段により選択された文書集合に含まれる文書の各々に対し、当該文書中におけるある単語の位置の前後の所定範囲に存在する単語列から、当該文書中におけるある単語の文脈の特徴量を表す、所定の構成の学習用の特徴量ベクトルを算出するための特徴量ベクトル算出手段と、文書集合選択手段により選択された文書集合に含まれる文書の各々に対して特徴量ベクトル算出手段により算出された学習用の特徴量ベクトルと、当該文書の検索時に使用された意味候補とを組にして学習用データを作成し、当該学習用データを用いた機械学習により、学習用の特徴量ベクトルと同じ構成の分類用の特徴量ベクトルが与えられると、当該分類用の特徴量ベクトルに対応する文脈中におけるある単語の意味として最適なものを、意味候補集合中から選択する所定の分類器を自動的に作成するための手段とを含む。

【００１５】

単語の文脈を、その単語の前後の所定範囲に存在する単語列から作成した学習用の特徴量ベクトルにより表す。こうした学習用の特徴量ベクトルを用いた機械学習により分類器を自動的に作成できる。その結果、適切な意味候補を人手を介在させることなく自動的に選択し、入力された単語の多義性を解消できる多義解消装置を提供できる。

【００１６】

より好ましくは、意味候補選択手段は、文書収集手段によって組合せの各々について収集された文書の集合のうち、集合に含まれる文書の数が多い所定の個数の集合を選択し、それら文書の集合に対応する意味候補のみを意味候補として選択する処理を行なうための手段を含む。

【００１７】

入力された単語と、ある意味候補との組合せに対して収集された文書の集合に含まれる文書の数が多いということは、その組合せを構成する単語が共起する可能性が高いということである。従ってそうした意味候補は入力された単語に対する適切な意味候補である可能性が高い。また、この時点で意味候補の上限個数が設定されるので、以後の処理を安定した時間で完了できる。その結果、適切な意味候補を、人手を介在させることなく自動的に、信頼性高く、安定した時間で選択し、入力された単語の多義性を解消できる多義解消装置を提供できる。

【００１８】

意味候補選択手段は、文書収集手段によって組合せの各々について収集された文書の集合のうち、集合に含まれる文書の数が予め定められるしきい値より大きな集合を選択し、それら文書の集合に対応する意味候補のみを意味候補として選択する処理を行なうための手段を含んでもよい。

【００１９】

入力された単語と、ある意味候補との組合せに対して収集された文書の集合に含まれる文書の数があるしきい値より多いということは、その組合せを構成する単語が共起する可能性が高いということである。従ってそうした意味候補は入力された単語に対する適切な意味候補である可能性が高い。その結果、適切な意味候補を、人手を介在させることなく自動的に、かつ信頼性高く選択し、入力された単語の多義性を解消できる多義解消装置を提供できる。

【００２０】

文書収集手段は、ある単語と、意味候補集合中の意味候補との組合せの各々について、インターネット上に存在するウェブページからなる仮想的コーパスから、当該組合せを構成する語が共起するウェブページの集合を検索し収集するための検索手段を含む。

【００２１】

インターネット上のウェブページは、多数の人間により作成され維持されている。従ってそこで使用されている単語の用法は非常に数多い使用例をカバーしている。そのため、そうした文書をもとに分類器を作成すると、分類結果の偏りをなくし、信頼性を高めることができる。

【００２２】

好ましくは、収集するための手段は、ある単語と、意味候補集合中の意味候補との組合せの各々について、インターネット上に存在するウェブページからなる仮想的コーパスから、当該組合せを構成する語が共起するウェブページの集合を検索し、所定の定数を上限とした要素数の集合として収集するための手段を含む。

【００２３】

一つの集合について収集されるウェブページの個数に上限が設けられる。そのため、分類器の学習が過大な負荷となるおそれは小さい。その結果、適切な意味候補を、人手を介在させることなく自動的に、かつ信頼性高く安定して選択し、入力された単語の多義性を解消できる多義解消装置を提供できる。

【００２４】

本発明の第２の局面に係るコンピュータプログラムは、コンピュータにより実行されると、当該コンピュータを、上記したいずれかの自然言語における多義解消装置として機能させるものである。

【発明を実施するための最良の形態】

【００２５】

以下、本発明の実施の形態について図を参照して説明する。実施の形態は三つある。第

１の実施の形態は、日本語の入力文に対する音声合成において、複数の仮名表記（読み）を持つ語の仮名表記を決定する装置に関する。第２の実施の形態は、英語のアクロニムに対し、英語の定義（フルスペル）を与える装置に関する。第３の実施の形態は、日本語から英語への翻訳において、日本語の単語に対し複数の英語の訳語が存在するときに、そのうちの一つを選択する装置に関する。すなわち、本発明において、ある単語の「意味」とは、日本語の場合に国語辞書にのっているような「意味」だけでなく、ある基準で見てその単語と等価であると評価できるような単語又は単語の集合又は文字列のことをいう。

【００２６】

なお、以下の実施の形態の説明に用いる図面において、同一の部品には同一の参照符号を付してある。それらの名称及び機能も同一である。従って、それらについての詳細な説明は繰返さない。なお、後述するように、各実施の形態は、コンピュータハードウェアと、その上で実行されるコンピュータプログラムとにより実現可能である。従って、以下に示すブロック図中の機能ブロックの一部については、それを実現するためのコンピュータプログラムのフローチャート形式でその機能及び構成を示す。

【００２７】

＜第１の実施の形態＞

［構成］

図１に、本発明の第１の実施の形態に係る音声合成システム３０のブロック図を示す。図１を参照して、音声合成システム３０は、音声合成の対象となる日本語の入力文を記憶するための入力文記憶部４０と、入力文記憶部４０から所定長を順次取出して記憶するための入力文バッファ４２と、日本語の単語と、その仮名表記とを対応付けて記憶した複数の辞書からなる辞書群４６と、入力文バッファ４２に含まれる文を形態素解析して、漢字を含む単語があれば辞書群４６を参照して仮名表記を検索し、仮名表記等の情報が付された形態素列を出力するための仮名変換部４４とを含む。

【００２８】

既に述べたように、漢字を含む単語の中には、複数の仮名表記を持つものがあり得る。音声合成のためには、それら複数の仮名表記の中で適切なものを選択する必要がある。音声合成システム３０は、そのために、仮名変換部４４及びいわゆるインターネット５２に接続され、仮名変換部４４がある単語Ｗについて複数の仮名表記候補Ｒｋ（ｋ＝１～Ｋ：Ｋは仮名表記候補の数）が存在することを検出したことに応答して、インターネット５２上でその単語Ｗと仮名表記候補Ｒｋとが共起するウェブページを、単語Ｗと仮名表記候補Ｒｋの組合わせの各々について検索し、得られたウェブページのテキストを学習データとした機械学習による分類によって、単語Ｗにふさわしい仮名表記を決定して仮名変換部４４に与えるための同形異音語解消処理部５０とを含む。すなわち、このシステムでは、インターネット５２上のウェブページの集合を、一つの仮想的なコーパスと見なして用例文書を収集している。

【００２９】

音声合成システム３０はさらに、仮名変換部４４が出力する、仮名表記付入力文を記憶するための仮名表記入力文記憶部５４と、音声合成のための、仮名表記に対応する音声を格納した音声データベース４８と、仮名表記入力文記憶部５４から仮名表記付入力文を読出し、音声データベース４８を参照して音声合成を行ない、アナログ音声信号を出力するための音声合成部５６と、音声合成部５６から出力されるアナログ音声信号を音声に変換するスピーカ５８とを含む。

【００３０】

本実施の形態では、同形異音語解消処理部５０がインターネット５２から検索するウェブページのテキストのうち、「スニペット」と呼ばれる部分を機械学習に用いる。「スニペット」とは、インターネットのいわゆる検索エンジンによる検索結果において、検索されたウェブページの内容を説明するための短文のことをいう。多くの場合、スニペットは、検索のキーワードとされた単語を含む部分のテキストからなる。なお、同形異音語解消処理部５０によるウェブページの検索には、独自の検索プログラムを用いてもよいが、本

実施の形態では、既存の検索サービスサイトを利用し、単語Wと仮名表記候補ＲｋとについてのＡＮＤ検索をするクエリを検索サービスサイトに対して発行し、その結果を得ることで行なっている。なお、本実施の形態では、処理時間を安定させるため、検索件数の上限として、一回の検索について１０００件という基準を設けている。

【００３１】

音声合成部５６による音声合成の部分は、本発明とは直接には関係しないため、その詳細についての説明はここでは行なわない。

【００３２】

図２に、同形異音語解消処理部５０の詳細なブロック図を示す。図２を参照して、同形異音語解消処理部５０は、単語Wが与えられると、入力文バッファ４２に記憶された入力文のうち、単語Wを中心とする所定長の窓に含まれる単語に基づいて行なう学習により、単語Wに関する所定の特徴ベクトルが与えられればその単語Wに対応する適切な仮名表記を出力するように学習可能な決定木８２と、仮名変換部４４から単語Wとその仮名表記候補Ｒｋとの組合せ（Ｗ，Ｒｋ）を受け、それらが共起するウェブページのスニペットをインターネット５２から収集し、その結果を用いて決定木８２の学習を行なうための決定木作成部８０と、仮名変換部４４に接続され、仮名変換部４４から、組合せ（Ｗ，Ｒｋ）中の単語Wと、入力文中における単語Wを中心とする所定範囲の単語列８５とが与えられると、それらから決定木８２による分類に適合した分類用特徴ベクトルを作成し、出力するための分類用特徴ベクトル作成部８４と、分類用特徴ベクトル作成部８４から出力される分類用特徴ベクトルを決定木８２に与え、その結果として決定木８２から得られる、分類結果である仮名表記を仮名変換部４４に与えるための分類実行部８６とを含む。

【００３３】

決定木作成部８０は、単語Wとその仮名表記Ｒｋとの組合せ（Ｗ，Ｒｋ）が与えられると、インターネット５２上でそれらが共起するウェブページを検索するための検索部１００と、検索部１００により検索されたウェブページのスニペットの集合（以下単に「ウェブページの集合」と呼ぶ。）を組合せ（Ｗ，Ｒｋ）ごとに記憶するための検索結果記憶部１０２と、組合せ（Ｗ，Ｒｋ）のうちで、取得されたウェブページの件数の降順にウェブページの集合をソートし、件数が上位であるＮ件（Ｎは自然数）のみを選択することにより、決定木８２のための学習データを作成するためのソート及び選択部１０４とを含む。本実施の形態では、このソート及び選択部１０４により選択された（Ｗ，Ｒｋ）に含まれるＮ個の仮名表記候補Ｒｋが、仮名表記候補として残され、後の決定木の学習に用いられる。

【００３４】

この処理では、他の文書集合は棄却され、それら文書集合の検索に用いられた仮名表記候補も棄却される。これは、単語Wと共起する頻度の低い仮名表記候補は候補として不適であると一般的に考えられるためである。もっとも、応用によってはそのように低頻度の仮名表記候補であっても棄却しない方がよい場合もあり得る。

【００３５】

決定木作成部８０はさらに、ソート及び選択部１０４により作成された学習データを記憶するための学習データ記憶部１０６と、検索対象となっている単語Wについて、学習データ記憶部１０６に記憶されている、その単語Wに関して検索された仮名表記候補Ｒｋのウェブページの各々から、所定の学習用特徴ベクトルを作成するための学習用特徴ベクトル作成部１０８と、学習用特徴ベクトル作成部１０８の作成した特徴ベクトルを記憶するための特徴ベクトル記憶部１１０と、特徴ベクトル記憶部１１０に記憶された特徴ベクトルを用いて決定木８２を学習させるための決定木学習部１１２とを含む。

【００３６】

図３は、図２に示す検索部１００を実現するためのコンピュータプログラムのフローチャートである。図３を参照して、このプログラムは、ある単語Wについての仮名表記の候補Ｒｋ（ｋ＝１～Ｋ）の各々について繰返されるステップ１３０～１３４の３つのステップを含む。

【００３７】
ステップ１３０では、クエリ「単語Ｗ　ａｎｄ　単語Ｒｋ」でウェブページを上限件数ＭＡＸ＝１０００件で検索する要求をインターネット上の検索エンジンに送信する。
【００３８】
ステップ１３２では、その検索結果として、単語Wと仮名表記候補Ｒｋ（ｋ＝１～Ｎ）とを含むスニペットの集合｛Ｓｎ（Ｗ，Ｒｋ）｝（ｎ＝１～Ｌｋ，ｋ＝１～Ｋ）を取得する。ただしここでＬｋは単語Wと仮名表記候補Ｒｋとの組合せに対して得られた検索結果の数である。
【００３９】
ステップ１３４では、各集合Ｓｎから仮名表記候補Ｒｋを削除することで、検索結果のスニペットの集合｛（Ｔｎ（Ｗ），Ｒｋ）｜ｎ＝１～Ｌｋ｝を作成する。
【００４０】
以上の３つのステップは、単語Wに対する仮名表記候補Ｒｋの全てに対して繰返される。
【００４１】
図２に示す検索部１００の機能はこのようなプログラムで実現される。
【００４２】
なお、ソート及び選択部１０４によって、検索件数が上位Ｎ個のスニペットの集合｛（Ｔｎ（Ｗ），Ｒｋ）｜ｎ＝１～Ｌｋ｝が抽出され、学習データ記憶部１０６に学習データとして記憶されるものとする。
【００４３】
図４に、図２に示す学習用特徴ベクトル作成部１０８の構成をブロック図形式で示す。図４を参照して、学習用特徴ベクトル作成部１０８は、学習データ記憶部１０６に記憶された学習データのスニペットの集合｛（Ｔｎ（Ｗ），Ｒｋ）｜ｎ＝１～Ｌｋ｝に含まれる各スニペットから、そのスニペット中に存在する単語Wをはさんで前後それぞれＭ個（合計２Ｍ個）の単語群（これら合計２Ｍ個の単語群を「窓」と呼ぶ。」）を抽出するための抽出部１５０と、学習データ記憶部１０６に記憶された学習データ１０６に出現する、単語W以外の語彙によって、決定木８２（図２参照）の学習に用いる分類用特徴ベクトルの構成を決定するためのベクトル構成決定部１５２と、ベクトル構成決定部１５２により決定された特徴ベクトルの構成に従い、抽出部１５０によりスニペットごとに抽出された単語群に基づいて各スニペットの特徴ベクトルの要素を算出して、各スニペットの特徴ベクトルを作成し、特徴ベクトル記憶部１１０に記憶させるための要素算出部１５４とを含む。
【００４４】
図５に、単語Wを中心とする「窓」の構成を模式的に示す。図５を参照して、学習用のスニペット１７０の単語列のうち、単語Wを中心としてその前後に存在する単語列を、単語Wを含めて、「W_{-m}，$W_{-(m-1)}$，$W_{-(m-2)}$，…，W_{-2}，W_{-1}，W、W_1，W_2，…，W_{m-2}，W_{m-1}，W_m」と書くことができる。単語Wを中心とし、その前のｍ個の単語からなる単語列１７４と、単語Wより後のｍ個の単語からなる単語列１７６とを含む単語列により、窓長２ｍの窓１７２が構成される。本実施の形態では、窓長を２Ｍとする。
【００４５】
ベクトル構成決定部１５２は、次のようにして特徴ベクトルの構成を決定する。すなわち、ベクトル構成決定部１５２は、学習データ記憶部１０６に存在する学習データ内に出現する単語の頻度を各単語について算出する。ベクトル構成決定部１５２はさらに、頻度が上位であるＨ個の単語のみを選択する。ベクトル構成決定部１５２はさらに、特徴ベクトルの次元をＨ次元とし、１番目～Ｈ番目の要素を、それぞれ頻度が１位～Ｈ位の単語に対応付ける。これにより特徴ベクトルの構成が決定される。この特徴ベクトルの要素数はＨ個である。各要素は０又は１の値をとる。各要素は、その要素に対応する単語がスニペット中の単語Wを中心とする窓長２Ｍの窓内に出現すると１の値となり、出現しないと０

の値となる。

【００４６】

従って、ある学習用のスニペットＴｉについて要素算出部１５４が行なう処理は次のような処理である。すなわち、要素算出部１５４は、このスニペットＴｉに対応するH次元の特徴ベクトルの各要素について、対応する単語がスニペットＴｉ中の、単語Wを中心とする窓長２Ｍの窓の中に出現するか否かを調べる。その要素の値は、その単語が出現すれば１、出現しなければ０となる。この処理をH個の要素の全てについて行なうことにより、スニペットＴｉの特徴ベクトルＶｉが算出される。この特徴ベクトルＶｉと、その特徴ベクトルが得られた組合せ（Ｗ，Ｒｋ）の仮名表記候補Ｒｋとを互いに関連付けて（特徴ベクトルに対する正解が仮名表記候補Ｒｋであるとして）決定木８２の学習に用いる。

【００４７】

図２に示す分類用特徴ベクトル作成部８４が行なう分類用の特徴ベクトルの作成も、基本的にはこれと同様である。すなわち、分類用特徴ベクトル作成部８４は、学習用特徴ベクトル作成部１０８のベクトル構成決定部１５２（図４参照）から、特徴ベクトルの各要素に対応する単語に関する情報を受け、処理対象となる単語Wについて、その単語Wを中心とする窓長２Ｍの窓内に所定の単語が出現するか否かによって、単語Wに対する分類用の特徴ベクトルを作成する。すなわち、この特徴ベクトルは、学習用特徴ベクトル作成部１０８によって作成される特徴ベクトルと全く同じ構成となる。

【００４８】

決定木学習部１１２は、機械学習によって決定木８２の学習を行なう。この学習方式については機械学習の分野で慣用されている事項であるので、ここではその詳細な説明は行なわない。

【００４９】

図６に、本実施の形態に係る要素算出部１５４により作成される決定木の一例である、「佐原」という単語に関する決定木２００を示す。図６を参照して、この決定木は、４つの中間のノード２１０，２１２，２１４及び２１６と、５つの終端のノード２３０，２３２，２３４，２３６及び２３８を含み、各ノード２１０，２１２，２１４及び２１６では、それぞれ窓内の単語が特定の条件を満たすか否かという質問がなされる。

【００５０】

ノード２１０の質問は、単語「佐原」を中心とする窓長２Ｍの窓内に、キーワード「千葉県」があるか、というものである。もしあればノード２３０に進み、「佐原」に対応する仮名表記として「さわら」が選択される。もしなければノード２１２に進む。なお、図６においては、「千葉県」のような具体的な単語について、窓内にあるか否かを聞いているが、実際の処理では、単語「佐原」の特徴ベクトル内において、単語「千葉県」に対応する要素（ビット）の値が１か０かを調べることによってこの判定を行なっている。

【００５１】

ノード２１２の質問は、キーワード「神奈川県」があるか、というものである。もしあればノード２３２に進み、「佐原」に対応する仮名表記として「さはら」が選択される。もしなければノード２１４に進む。

【００５２】

ノード２１４の質問は、キーワード「成田」があるか、というものである。もしあればノード２３４に進み、「佐原」に対応する読みとして「さわら」２３４が選択される。もしなければノード２１６に進む。

【００５３】

ノード２１６の質問は、キーワード「横須賀」があるか、というものである。もしあればノード２３６に進み、「佐原」に対応する仮名表記として「さはら」が選択される。もしなければノード２３８に進み、「佐原」に対応する仮名表記として「さわら」が選択される。

【００５４】

本実施の形態では、基本的に各単語に対し、決定木２００が作成される。ある単語に対

応する特徴ベクトルが与えられると、その単語に対応する決定木を特徴ベクトルの各要素の値に従ってたどることにより、その単語の仮名表記が選択される。

【００５５】

［動作］

図１～図６を参照して、上記した音声合成システム３０は以下のように動作する。図１に示す入力文記憶部４０には、音声合成の対象となる日本語の文が予め記憶される。そのうちの所定長部分が読出され、入力文バッファ４２に記憶される。

【００５６】

仮名変換部４４は、入力文バッファ４２に記憶された文について辞書群４６を参照して形態素解析を行なう。その結果、各単語の品詞、仮名表記（漢字の場合）、活用型、活用形などが決定される。もしも一つの単語について複数の仮名表記が得られた場合（すなわち同形異音語が存在する場合）、仮名変換部４４は、その単語（単語Wとする。）と、仮名表記の組合せをそれぞれ同形異音語解消処理部５０に与える。以下の説明では、構成のときに使用した表記を用いる。すなわち、ある単語Wに対して得られたK個の仮名表記候補を仮名表記候補R_1～R_Kとする。

【００５７】

図２を参照して、検索部１００は、単語Wと、仮名表記候補Ｒｋ（ｋ＝１～Ｋ）との組合せ（Ｗ，Ｒｋ）が与えられると、（単語Ｗ　ａｎｄ　単語Ｒｋ）をクエリとしてインターネット５２上の検索エンジンに検索件数上限＝１０００件という条件で検索要求を送信する（図３のステップ１３０）。そして、この検索要求に応答して検索エンジンから得られたウェブページのスニペットの集合｛Ｓｎ（Ｗ，Ｒｋ）｝（ｎ＝１～Ｌｋ）を取得する（図３のステップ１３２）。ここでＬｋはクエリ（単語Ｗ　ａｎｄ　単語Ｒｋ）に対して得られた検索結果（ウェブページ）の数である。このスニペットの集合の各々から単語Ｒｋを削除して得られた検索結果のスニペットの集合が検索結果記憶部１０２に記憶される（図３のステップ１３４）。これらスニペットの集合は、（単語Ｗ，仮名表記候補Ｒｋ）の組合せごとに得られる。スニペットの集合の各々の要素の数Ｌｋの上限ＭＡＸは、本実施の形態では、上記したように１０００である。

【００５８】

検索部１００は、単語Wと仮名表記候補Ｒｋとの組合せの各々に対し、上記した処理を実行する。すなわち、図３におけるステップ１３０～１３４の処理を各組合せに対し実行する。その結果、検索結果記憶部１０２には、これら組合せの各々について、検索結果のスニペットの集合｛Ｓｎ（Ｗ，Ｒｋ）｝が記憶される。

【００５９】

ソート及び選択部１０４は、検索結果記憶部１０２に記憶されたスニペットの集合｛Ｓｎ（Ｗ，Ｒｋ）｝を、その要素の数Ｌｋをキーに降順にソートする。ソート及び選択部１０４はさらに、ソート結果のうち、上位Ｎ個のスニペットの集合｛（Ｔｎ（Ｗ），Ｒｋ）｜ｎ＝１～Ｌｋ｝を選択して、それらスニペットが得られた仮名表記要素Ｒｋと関連付けて学習データ記憶部１０６に学習データとして記憶させる。すなわち学習データ記憶部１０６には、スニペットの集合のうち、検索結果の多かったものから順番にＮ個が記憶される。

【００６０】

図４を参照して、学習用特徴ベクトル作成部１０８のベクトル構成決定部１５２は、学習データ記憶部１０６に学習データが記憶されると、これら学習データに出現する単語の頻度を各単語について算出する。ベクトル構成決定部１５２はさらに、出現頻度が上位Ｈ番目までの単語を選択する。特徴ベクトルの１番目～Ｈ番目の要素を出現頻度１位～Ｈ位の単語に対応付けることにより、特徴ベクトルの構成が決定される。ベクトル構成決定部１５２は、この特徴ベクトルの構成（すなわち特徴ベクトルの各要素に対応する単語に関する情報）を図２に示す分類用特徴ベクトル作成部８４及び図４に示す要素算出部１５４に与える。

【００６１】

一方、抽出部１５０は、学習データ記憶部１０６に記憶されている各スニペットについて、単語Ｗを中心とする窓長２Ｍの窓を抽出して要素算出部１５４に与える。

【００６２】

要素算出部１５４は、ベクトル構成決定部１５２から与えられるベクトル構成に従い、抽出部１５０から与えられる窓に含まれる単語に基づいて、各スニペットの特徴ベクトルの各要素の値を算出する。その結果、各スニペットの特徴ベクトルが得られる。要素算出部１５４は、各スニペットを、そのスニペットが検索されたときの仮名表記候補Ｒｋと関連付けて特徴ベクトル記憶部１１０に学習用データとして記憶させる。

【００６３】

図２を参照して、決定木学習部１１２は、特徴ベクトル記憶部１１０に記憶された特徴ベクトルと、それら特徴ベクトルに関連付けられた仮名表記候補とを用いた機械学習により、決定木８２の学習を行なう。

【００６４】

以上の処理によって、決定木８２は、ある単語Ｗを中心とする窓長２Ｍの窓中の単語列、すなわち単語Ｗの文脈、を表す特徴ベクトルが与えられると、その文脈における単語Ｗの仮名表記として最適なものを出力するように機能するようになる。

【００６５】

一方、仮名変換部４４は、分類用特徴ベクトル作成部８４に対し、同形異音語の解消を要求する単語Ｗと、入力文において単語Ｗを中心とする窓長２Ｍの窓に含まれる単語列８５とを与える。分類用特徴ベクトル作成部８４は、単語Ｗについて、仮名変換部４４より与えられた、入力文中のその単語Ｗを中心とする窓長２Ｍの窓に含まれる単語列８５と、図４に示すベクトル構成決定部１５２から与えられたベクトル構成とによって、要素算出部１５４と同様の処理により単語Ｗの特徴ベクトルを作成し、分類実行部８６に与える。

【００６６】

分類実行部８６は、この特徴ベクトルを決定木８２に与える。決定木８２は、単語Ｗを中心とする窓長Ｍから上記方法によって作成した特徴ベクトルが与えられると、単語Ｗの仮名表記として適切なものを出力するように学習済みである。分類実行部８６は、この仮名表記を決定木８２から得て、仮名変換部４４に与える。

【００６７】

仮名変換部４４は、このようにして同形異音語解消処理部５０から得られた仮名表記を、問題となった単語Ｗに形態素分析の結果と同様にして付加する。仮名変換部４４はさらに、形態素解析が終わり、品詞、仮名表記（漢字の場合）、活用型、活用形などの情報が付された形態素列を音声合成部５６に与える。この場合、同形異音語については既に同形異音語解消処理部５０により解消されているため、一つの単語には一つの仮名表記しか付されていない。

【００６８】

音声合成部５６は、与えられた形態素列に基づき、形態素に付された仮名表記などを用いて音声データベース４８から適切な音声波形を抽出し、波形接続処理によって合成音声波形データを作成し、さらにこの合成音声波形データをアナログ変換してスピーカ５８に与える。スピーカ５８はこの音声信号を音声に変換する。

【００６９】

以上のように音声合成システム３０によれば、入力文記憶部４０に記憶された入力文に同形異音語が含まれていても、同形異音語解消処理部５０によって同形異音語が解消され、一つの仮名表記のみがその単語に割当てられる。インターネット５２上のウェブページをいわば仮想的なコーパスとして用い、自動的にこの同形異音語の解消のための決定木の学習が行なわれる。人手で学習データを作成する必要がなく、同形異音語の解消のための手間を従来と比較してはるかに少なくできる。さらに、インターネット５２上で検索されるウェブページは多数の人により作成されたものであるため、少数の人が学習データを作成する場合と比較して、学習データの偏りが少なく、そのカバーする範囲も広くなる。従って、同形異音語の解消の信頼性が従来より高くなるという効果がある。

【００７０】
［コンピュータによる実現］
上記した第１の実施の形態に係る音声合成システム３０は、既に述べたようにコンピュータハードウエア及び当該コンピュータハードウエア上で実行されるコンピュータソフトウエアにより実現される。図７に音声合成システム３０を実現するための一般的なコンピュータシステム２５０の外観を示し、図８にこのコンピュータシステム２５０の内部構成をブロック図形式で示す。
【００７１】
図７を参照して、コンピュータシステム２５０は、コンピュータ２６０と、いずれもコンピュータ２６０に接続されるモニタ２６２、キーボード２６６、マウス２６８、マイクロホン２９０及び一対のスピーカ５８とを含む。コンピュータ２６０には、ＤＶＤ（Ｄｉｇｉｔａｌ　Ｖｅｒｓａｔｉｌｅ　Ｄｉｓｃ）の再生及び記録が可能なＤＶＤドライブ２７０と、所定の規格に従った半導体メモリ記憶装置が装着可能なメモリポート２７２とが備えられている。コンピュータ２６０の内部構成については図８を参照して後述する。
【００７２】
図８を参照して、コンピュータ２６０は、図７に示すＤＶＤドライブ２７０及びメモリポート２７２に加え、ＣＰＵ（中央演算処理装置）２７６と、ＣＰＵ２７６に接続されたバス２８６と、いずれもバス２８６に接続されたＲＯＭ（読出専用メモリ）２７８、ＲＡＭ（ランダムアクセスメモリ）２８０、ハードディスク２７４、ネットワークインタフェース２９６、及びサウンドボード２８８を含む。
【００７３】
ＤＶＤドライブ２７０には、ＤＶＤ２８２が装着される。メモリポート２７２には半導体メモリ記憶装置２８４が装着される。ＣＰＵ２７６は、バス２８６並びにＤＶＤドライブ２７０及びメモリポート２７２をそれぞれ介して、ＤＶＤ２８２及びメモリ２８４をアクセスできる。
【００７４】
キーボード２６６、マウス２６８、モニタ２６２はいずれも図示しないインタフェースを介してコンピュータ２６０のバス２８６に接続される。スピーカ５８及びマイクロホン２９０は、サウンドボード２８８に接続される。このコンピュータシステム２５０において、ＣＰＵ２７６で実行される音声合成プログラムは、最終的にはデジタル形式の音声波形データを生成する。サウンドボード２８８はその音声波形データをＣＰＵ２７６から受取ると、アナログ信号に変換してスピーカ５８を介して音声を発生させる処理をする。
【００７５】
上記実施の形態における入力文記憶部４０、辞書群４６、仮名表記入力文記憶部５４、音声データベース４８、検索結果記憶部１０２、学習データ記憶部１０６、特徴ベクトル記憶部１１０等は、ＲＡＭ２８０、ハードディスク２７４、ＤＶＤディスク２８２、半導体メモリ記憶装置２８４のいずれでも実現できる。実際には、格納するデータの容量、読出し、書込みに要求される速度などによって、最も効率のよい記憶装置が各記憶部を実現するために選択される。
【００７６】
上記した第１の実施の形態に係る音声合成システム３０を実現するためのコンピュータプログラムは、単一のプログラムでもよいし、複数のプログラムを組合せたものでもよい。特に、上記した各部の機能のうち、図１に示す仮名変換部４４において行なわれる形態素解析処理、音声合成部５６において行なわれる音声合成処理、図２に示す検索部１００が実行するスニペットの検索処理、ソート及び選択部１０４が実行するソート及び選択処理、決定木学習部１１２が実行する決定木８２の学習処理などについては、既に広く流布しているプログラムをそのまま使用できる。もちろん、これらプログラムは汎用に作成されているため、適切な調整を行なうことは要求されるが、それらはこの技術分野における通常の知識を持つ者にとっては、目的に照らして容易に実現できる範囲に留まる。
【００７７】

さらに、学習用特徴ベクトル作成部１０８、分類用特徴ベクトル作成部８４での処理についても、上記した説明に基づいて、当該技術分野の通常の知識を持つものであれば、仕様に応じて適宜実現することが可能である。

【００７８】

これらプログラムは、例えばＤＶＤディスク２８２等のような記憶媒体に記憶され、又はインターネット５２等のネットワークを通じて流通し、通常はハードディスク２７４等の不揮発外部記憶装置に記憶される。そして実行時にはハードディスク２７４からＲＡＭ２８０にコピーされ、ＣＰＵ２７６内の図示しないプログラムカウンタにより指し示されるアドレスから読出された命令がＣＰＵ２７６により実行され、上記した所期の機能を実現する。コンピュータハードウェアそのものの動作形態については周知であるので、ここではこれ以上の詳細な説明は行なわない。

【００７９】

＜第２の実施形態＞

図９に、本発明の第２の実施の形態に係る、複数の定義を有する英語のアクロニムに対し、適切な定義を与える多義アクロニム解消システム３３０の構成をブロック図形式で示す。この多義アクロニム解消システム３３０は、アクロニムの近傍に、そのアクロニムの定義を与えている文書が多いこと、アクロニムの近傍に存在する単語は、その文書の分野によって何らかの傾向を持っていることを利用して、実施の形態１における同形異音語の解消と同じ原理によって、アクロニムに適切な定義を与えるものである。

【００８０】

図９を参照して、この多義アクロニム解消システム３３０は、アクロニムを含む可能性のある入力文を記憶するための入力文記憶部３４０と、入力文記憶部３４０に記憶された入力文の所定部分を読込むための入力文バッファ３４２と、アクロニム及びその定義のリストよりなるデータからなる辞書群３４６と、入力文バッファ３４２に格納された入力文を形態素解析し、定義が付されていないアクロニムを見出すと、辞書群３４６によって当該アクロニムの定義を決定し、入力文中の当該アクロニムに当該定義を付して入力文を出力するためのアクロニム解釈部３４４とを含む。

【００８１】

多義アクロニム解消システム３３０はさらに、アクロニム解釈部３４４から出力される、アクロニムに定義が付された入力文を記憶するためのアクロニム定義付入力文記憶部３５４と、アクロニム定義付入力文記憶部３５４に記憶された入力文の意味を理解するための文章理解装置３５６とを含む。

【００８２】

既に述べたように、アクロニムの中には複数の定義を持つものもあり得る。そうした場合に、アクロニム解釈部３４４がアクロニムに複数の定義を付して出力することはできない。そうすると、文章理解装置３５６における文章理解の障害となるからである。従って、入力文中で定義されていないアクロニムに対し、複数の定義が辞書群３４６から見出された場合、何らかの手段によりそれらの中の適切な一つを自動的に選択できるようにする必要がある。

【００８３】

こうした問題を解決するために、本実施の形態に係る多義アクロニム解消システム３３０は、アクロニム解釈部３４４及びインターネット５２に接続され、アクロニム解釈部３４４から、アクロニムと、そのアクロニムに対して得られた複数の定義候補と、アクロニムの前後の所定の窓中に存在する単語列とが与えられると、インターネット５２をコーパスとして用いた学習処理により、与えられた複数の定義候補のうち、与えられた単語列に対して最も適切と思われるものを選択し、アクロニム解釈部３４４に与える処理を行なうための多義アクロニム解消処理部３５０を含む。

【００８４】

多義アクロニム解消処理部３５０の構成の詳細についてはここでは述べないが、多義アクロニム解消処理部３５０の構成及び動作は第１の実施の形態における同形異音語解消処

理部５０と同様である。すなわち多義アクロニム解消処理部３５０は、以下の手順でアクロニムに対する適切な定義を決定する。

【００８５】

（１）アクロニムＡと定義候補Ｄｋ（ｋ＝１～Ｋ：Ｋは定義候補の数）が与えられると、定義候補Ｄｋの各々について、アクロニムＡと定義候補Ｄｋとが共起するウェブページのスニペットに対する検索要求をインターネット５２上の検索エンジンに与える。

【００８６】

（２）検索結果として、アクロニムＡと定義候補Ｄｋとを含むスニペットの集合｛Ｓｎ（Ａ，Ｄｋ）｝（ｎ＝１～Ｌｋ）（ただしＬｊ（ｊ＝１～ｋ）はアクロニムＡと定義候補Ｄｋとの組合せに対して検索されたスニペットの数を表す。）を取得する。

【００８７】

（３）このスニペットの集合｛Ｓｎ（Ａ，Ｄｋ）｝の各々から、定義候補Ｄｋを削除することによって、検索結果のスニペットの集合｛（Ｔｎ（Ａ），Ｄｋ）｜ｎ＝１～Ｌｋ｝を作成する。

【００８８】

（４）上記した３つの処理を、全ての定義候補Ｄｋに対して繰返す。

【００８９】

（５）検索されたウェブページのスニペットの集合Ｓｎを、それらに含まれるウェブページの数（検索結果の数）の降順でソートし、さらにその内で上位Ｎ個のみを選択することで、Ｎ個の学習用のスニペットの集合｛（Ｔｎ（Ａ），Ｄｋ）｜ｎ＝１～Ｌｋ｝が抽出され、学習データとして記憶される。

【００９０】

（６）この学習データを用い、図４に示す学習用特徴ベクトル作成部１０８と全く同様にして学習用の複数個の特徴ベクトルが作成される。特徴ベクトルの作成の仕方も第１の実施の形態の場合と全く同様である。特徴ベクトルの作成時の窓長も第１の実施の形態と同様、２Ｍと表すことにする。

【００９１】

（７）これらの特徴ベクトルと、それら特徴ベクトルを与えたスニペットが検索されたときの検索に用いられた定義候補とを関連付けて学習用のデータとする。

【００９２】

（８）この学習用のデータを用い、決定木の学習を行なう。この学習の結果、決定木は、入力文のうち、多義解消の対象となるアクロニムＡを中心とする窓長２Ｍに含まれる単語により作成される特徴ベクトルが与えられると、そのアクロニムに対する適切な定義を出力するようになる。

【００９３】

（９）入力文の中の、多義解消の対象となるアクロニムＡを中心とし、窓長２Ｍの窓から決定木のための特徴ベクトルを作成する。

【００９４】

（１０）この特徴ベクトルを決定木に与えることにより、決定木からはアクロニムＡの定義を一つだけ選択する出力が得られる。この出力を多義アクロニム解消処理部３５０からアクロニム解釈部３４４に与えることにより、アクロニム解釈部３４４は当該アクロニムに対し、多義アクロニム解消処理部３５０から与えられたただ一つの定義を付して、アクロニム定義付入力文記憶部３５４に出力できる。

【００９５】

＜第３の実施の形態＞

図１０に、第３の実施の形態に係る日本語－英語の自動翻訳システム４３０のブロック図を示す。図１０を参照して、この自動翻訳システム４３０は、日本語の入力文を記憶するための日本文記憶部４４０と、日本文記憶部４４０に記憶された日本文の所定量を記憶するための入力文バッファ４４２と、日本語から英語への１又は複数の辞書からなる辞書群４４６と、自動翻訳の前処理として、入力文バッファ４４２に記憶された日本文を形態

素解析し、各単語について辞書群４４６を参照して英語の訳語を割当て、出力するための訳語決定部４４４と、このように前処理された訳語付日本文を記憶するための訳語付日本文記憶部４５４と、訳語付日本文記憶部４５４に記憶された訳語付日本文を、その訳語を使用しながら英語に翻訳する自動翻訳装置４５６とを含む。

【００９６】

しかし、既に述べたとおり、入力される一つの日本語単語に複数の英語の訳語候補が存在する場合があり得る。そうしたときにそれら複数の英語の訳語候補を日本語単語にそのまま付して訳語決定部４４４から出力すると、自動翻訳装置４５６における翻訳に支障が生ずる。そのために、何らかの手段でこれら複数の訳語候補の中から適切なものを選択する必要がある。

【００９７】

そのために、本実施の形態に係る自動翻訳システム４３０は、訳語決定部４４４及びインターネット５２に接続され、訳語決定部４４４から、日本語の単語と、その単語に対して得られた複数の訳語候補と、入力文における当該日本語の単語の前後の所定の窓中に存在する単語列とが与えられると、インターネット５２をコーパスとして用いた学習処理により、与えられた複数の訳語候補のうち、与えられた単語列に対して最も適切と思われるものを選択し、訳語決定部４４４に与える処理を行なうための多義訳語解消処理部４５０を含む。

【００９８】

多義訳語解消処理部４５０の構成の詳細についてはここでは述べないが、多義訳語解消処理部４５０における処理が、第１の実施の形態の同形異音語解消処理部５０における処理、及び第２の実施の形態の多義アクロニム解消処理部３５０における処理と同一であり、従ってその構成も同形異音語解消処理部５０の構成と同一であることが理解されるであろう。

【００９９】

訳語決定部４４４は、入力文バッファ４４２中の文を読出し、形態素解析して、各単語について辞書群４４６を参照することにより英語の訳語を割当て、訳語付日本文記憶部４５４に出力していく。複数の訳語候補が一つの日本語単語について出現した場合、訳語決定部４４４はその日本語単語と、複数の訳語候補とを多義訳語解消処理部４５０に引渡し、多義性の解消を依頼する。多義訳語解消処理部４５０は、第１の実施の形態における同形異音語解消処理部５０と全く同じ動作によって決定木を作成し、入力文のうち、与えら得た日本語単語の前後の窓内の単語列を用いて特徴ベクトルを作成し、決定木に与えることにより適切な訳語候補を得て、訳語決定部４４４に返す。訳語決定部４４４は問題となった日本語単語に、多義訳語解消処理部４５０から与えられたただ一つの訳語を付し、訳語付日本文記憶部４５４に出力する。従って、自動翻訳装置４５６における自動翻訳処理に支障が生ずることはない。

【０１００】

以上、第１～第３の実施の形態の説明から明らかなように、本発明に係る多義性の解消、又はあいまい性の解消は、自然言語処理の分野の広い範囲にわたり、容易に適用できる。しかも、多義性の解消を行なう部分の仕組みは基本的に同一でよい。もちろん、解消処理の細部において様々な設計事項はあり得るが、ある分野で有効な方式は、基本的にそのままの形で他の分野についても適用可能である。

【０１０１】

例えば日本語と英語との間の翻訳のみならず、あらゆる言語の間の単語の翻訳に、言語の相違にかかわらず本発明に係る多義性又はあいまい性の解消をする装置を適用できる。第１の実施の形態における同形異音語の解消を行なう機構も、言語にかかわらずほとんどそのまま適用できる。もちろん、言語に特有の調整が必要な場合もあり得るが（例えば日本語における形態素解析）、その部分は自然言語処理での前提として必ず前もって行なわれているとすれば、多義性又はあいまい性の部分の仕組みは言語に係らず同一でよい。

【０１０２】

従って、自然言語処理の分野の広い領域において、本発明を適用することができ、しかもある領域から別の領域への移植も極めて簡単に実現できる。

【０１０３】

＜可能な変形例＞

上記した実施の形態では、適切な仮名表記、アクロニムの定義、及び訳語を決定するために、決定木を用いた。しかし本発明は決定木を用いるものには限定されず、インターネットから収集した学習データによって、対象となる単語又は単語列がおかれた文脈（環境）によって、目的物として複数のうちからどれを選択するかを機械学習により学習できるものであれば、どのような分類方法でも用いることができる。例えば、ナイーブベイズ、決定リスト、ｋ－最近隣法、オンラインアルゴリズム、最大エントロピー法、サポートベクトルマシン、ブースティングなどを利用できる。

【０１０４】

また、上記した実施の形態では学習データとしてウェブページのスニペットを収集したが、本発明がそのような実施の形態に限定されないことはもちろんである。例えばウェブページ全体を処理の対象としてもよい。また、例えば一つの単語Ｗと仮名表記Ｒｋとの組み合わせに対して収集するウェブページの数の上限ＭＡＸを１０００に限定しているが、この数が自由に変更できることはいうまでもない。また、このような限定を用いないことも可能である。

【０１０５】

さらに、上記した実施の形態では、問題となる単語と、その単語と対となるべきいくつかの候補が与えられると、その時点でインターネットにアクセスし、決定木を作成している。しかし本発明はそのような実施の形態には限定されない。例えば、予め何らかのテスト文に対し、上記したような処理をすることにより、テスト文中に含まれる、何らかのあいまい性を持ついくつかの単語について、そのあいまい性を解消するための分類装置を予め準備しておいてもよい。そうした分類装置を多数の単語に対して一つずつ予め準備しておけば、その単語が与えられてから分類装置の学習を行なったりする必要はなく、直ちに適切な答えを与えることができる。もしもそれら複数の分類装置ではあいまい性が解消できない単語であれば、そのときに上記実施の形態で示したように改めて一つの分類装置を作成して適切な答えを得るようにすればよい。

【０１０６】

また、上記した第１の実施の形態では、ソート及び選択部１０４により選択される仮名表記候補は、検索部１００によりヒットしたウェブページの数の多い上位Ｎ件（Ｎは複数）であった。第２の実施の形態及び第３の実施の形態の場合も同様である。しかし本発明はそのような実施の形態には限定されない。例えば、ソート及び選択部１０４の処理でヒット数の多かった最上位の１件の仮名表記候補のみを単語Ｗの仮名表記として採用してもよい。この場合には、決定木は１：１の分類を行なうものとして機能する。もっとも、この方法では単語Ｗの文脈が全く考慮されないので、結果の信頼性は低く、あいまい性の解消とはいえない。

【０１０７】

また、第１の実施の形態のソート及び選択部１０４の処理で、ヒットしたウェブページの数の多い上位Ｎ件ではなく、所定のしきい値以上のウェブページがヒットしたものを全て仮名表記候補として選択してもよい。又は、全ヒット数のうち、上位から各候補の割合を積算し、所定割合を超えるまでのものを、その数にかかわらず全て仮名表記候補として採用してもよい。

【０１０８】

さらに、上記実施の形態では、一つの単語を単位としてその意味候補を決定している。しかし本発明はそのような実施の形態には限定されない。意味候補の集合を作成するための辞書の見出しとして、例えば複数の単語からなる句を設けておくことにより、その句の意味についても、複数の意味集合の中から適切なものを選択できるようになる。

【０１０９】

そして、そのようにして得られた分類装置を随時蓄積しておくことにより、直ちにあいまい性を解消できる単語が増加することになり、好ましい。

【０１１０】

今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、特許請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内でのすべての変更を含む。

【図面の簡単な説明】

【０１１１】

【図１】本発明の第１の実施の形態に係る音声合成システム３０のブロック図である。

【図２】図１に示す同形異音語解消処理部５０のブロック図である。

【図３】図２の検索部１００を実現するためのコンピュータプログラムの制御構造を示すフローチャートである。

【図４】図２の学習用特徴ベクトル作成部１０８のブロック図である。

【図５】「窓」の概念について説明するための図である。

【図６】決定木の一例を模式的に示す図である。

【図７】第１の実施の形態に係る音声合成システム３０を実現するコンピュータシステム２５０の外観を示す図である。

【図８】図７に示すコンピュータシステム２５０の内部構成を示すブロック図である。

【図９】本発明の第２の実施の形態に係る多義アクロニム解消システム３３０のブロック図である。

【図１０】本発明の第３の実施の形態に係る自動翻訳システム４３０のブロック図である。

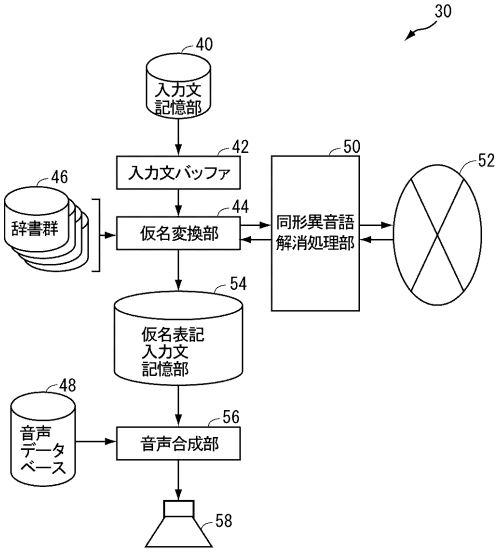
【符号の説明】

【０１１２】

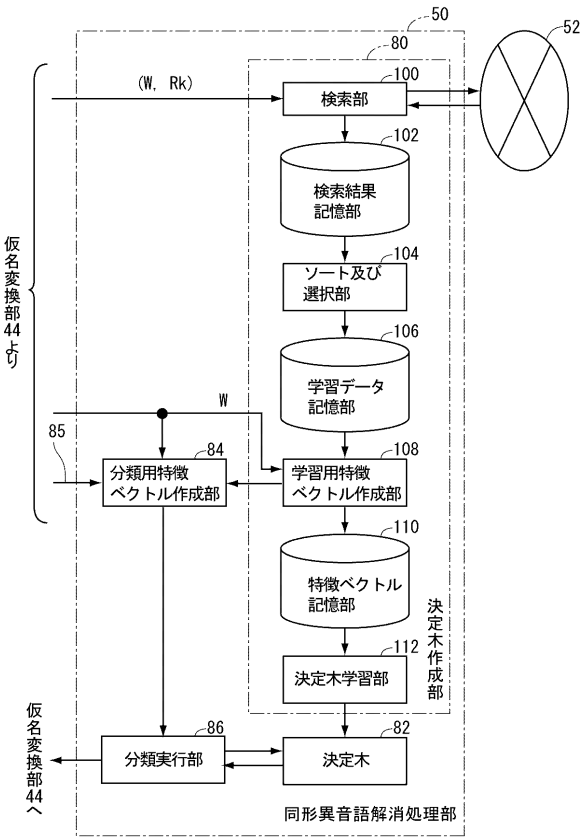
３０　音声合成システム
４０，３４０　入力文記憶部
４２，３４２，４４２　入力文バッファ
４４　仮名変換部
４６，３４６，４４６　辞書群
４８　音声データベース
５０　同形異音語解消処理部
５２　インターネット
５４　仮名表記入力文記憶部
５６　音声合成部
５８　スピーカ
８０　決定木作成部
８２　決定木
８４　分類用特徴ベクトル作成部
８６　分類実行部
１００　検索部
１０２　検索結果記憶部
１０４　ソート及び選択部
１０６　学習データ記憶部
１０８　学習用特徴ベクトル作成部
１１０　特徴ベクトル記憶部
１１２　決定木学習部
３４４　アクロニム解釈部
３５０　多義アクロニム解消処理部
３５４　アクロニム定義付入力文記憶部

４４０　日本文記憶部
４４４　訳語決定部
４５０　多義訳語解消処理部
４５４　訳語付日本文記憶部
４５６　自動翻訳装置

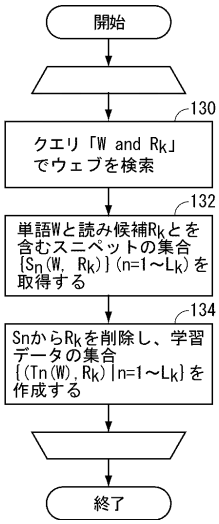
【図１】



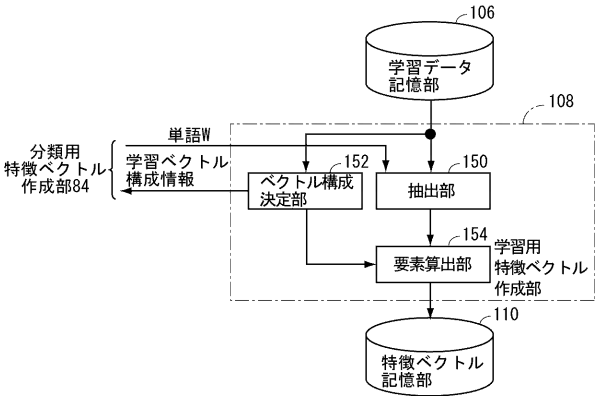
【図２】



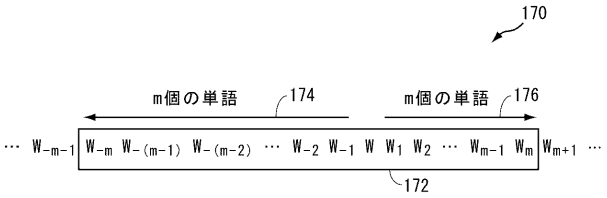
【図３】



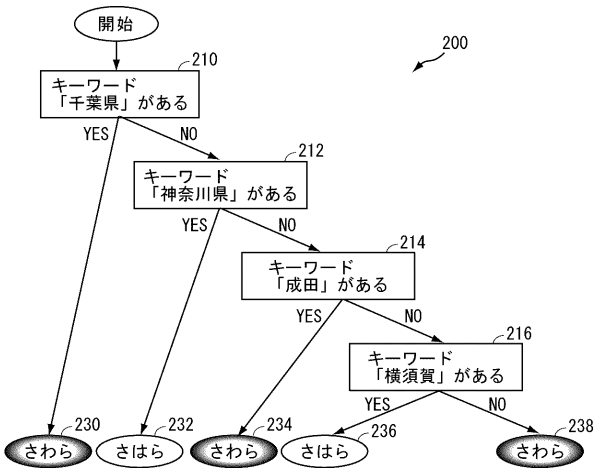
【図４】



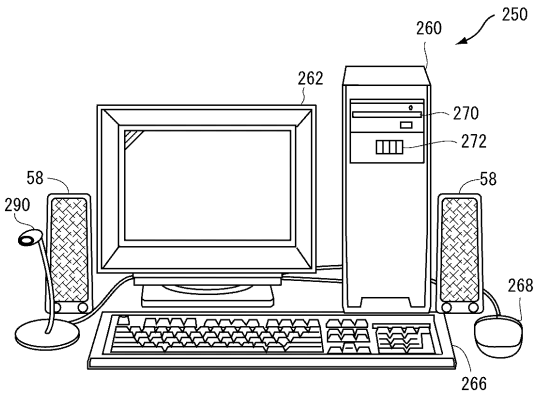
【図５】



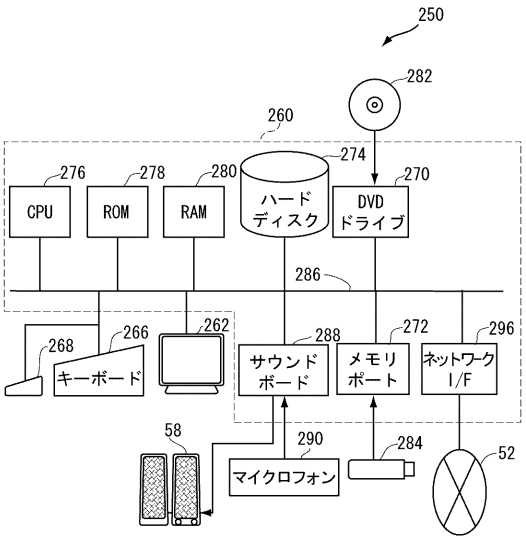
【図６】



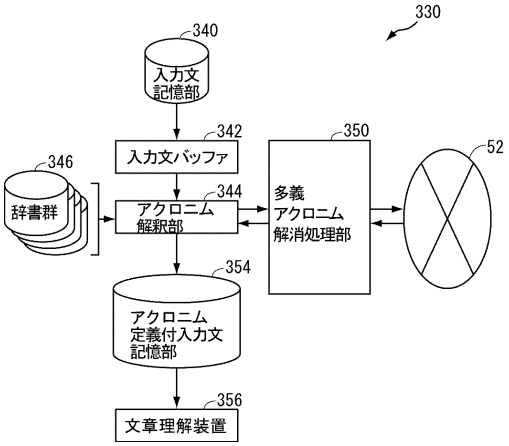
【図７】



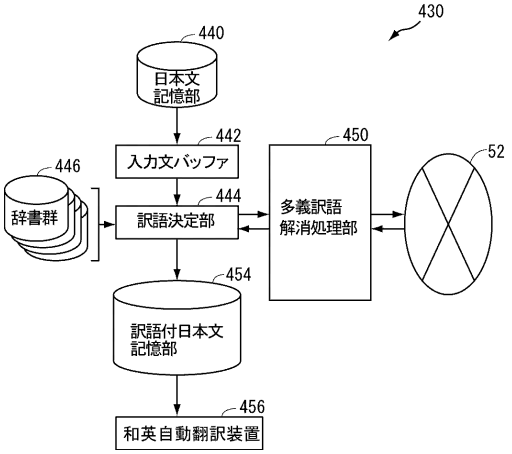
【図８】



【図９】


330
340
入力文
記憶部
342
入力文バッファ
346
辞書群
344
アクロニム
解釈部
350
多義
アクロニム
解消処理部
52
354
アクロニム
定義付入力文
記憶部
356
文章理解装置

【図１０】


430
440
日本文
記憶部
442
入力文バッファ
446
辞書群
444
訳語決定部
450
多義訳語
解消処理部
52
454
訳語付日本文
記憶部
456
和英自動翻訳装置

フロントページの続き

審査官　成瀬　博之

(56)参考文献　村田真樹　他４名，種々の機械学習手法を用いた多義解消実験，電子情報通信学会技術研究報告，日本，社団法人電子情報通信学会，２００１年　5月１１日，第101巻第40号(NLC2001 1-8)，7-14頁
前田亮 他２名，言語横断情報検索におけるＷｅｂ文書群による訳語曖昧性解消，情報処理学会論文誌，日本，社団法人情報処理学会，２０００年１０月１５日，第41巻No.SIG6(TOD7)，12-21頁
藤井敦，コーパスに基づく多義性解消，人工知能学会誌，日本，社団法人人工知能学会，１９９８年１１月　1日，第13巻第6号，904-911頁

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｆ　１７／２０－１７／２８