(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2003－122750

（Ｐ２００３－１２２７５０Ａ）

(43)公開日　平成15年４月25日(2003.4.25)

(51)Int.Cl.[7]	識別記号	ＦＩ	ﾃｰﾏｺｰﾄﾞ (参考)
G06F 17/27		G06F 17/27	M 5B091

審査請求　有　請求項の数９　ＯＬ　（全16頁）

(21)出願番号　特願2001－311453(Ｐ2001－311453)

(22)出願日　平成13年10月９日(2001.10.9)

特許法第30条第１項適用申請有り　平成13年７月10日　社団法人電子情報通信学会発行の「電子情報通信学会技術研究報告　信学技報　Ｖｏｌ.101　Ｎｏ.190」に発表

(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

(74)代理人　100119161
弁理士　重久　啓子　（外１名）

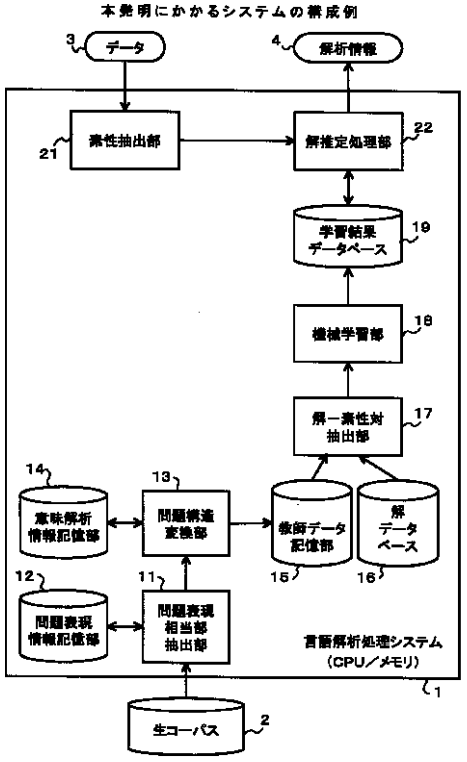
Ｆターム(参考)　5B091 AA15 CA12 CA14 CC01 EA01

(54)【発明の名称】機械学習法を用いた言語解析処理システム、教師データ生成処理方法、機械学習法を用いた言語解析処理方法、機械学習法を用いた言語省略解析処理システム

(57)【要約】

【課題】　言語解析において、広範かつ多数の情報を教師信号として用いることができる機械学習法を用いた言語解析処理システムを実現する。

【解決手段】　問題表現相当部抽出部11は問題表現情報記憶部12を参照して解析情報が付与されていない生コーパス2 から問題表現に相当する部分を抽出し、問題構造変換部13は、当該抽出部分を問題表現に変換して抽出した解と教師データを生成する。解－素性対抽出部17は教師データ記憶部15に保存された教師データから解と素性の集合の組を抽出し、機械学習部18は抽出した組からどのような素性のときにどのような解となりやすいかを学習した結果を保存する。素性抽出部21は入力されたデータ3 から素性の集合を抽出し、解推定処理部22は学習結果データベース19をもとに素性の集合からその素性の場合にどのような解になりやすいかを推定した解析情報 4 を出力する。



1

【特許請求の範囲】

【請求項１】　機械学習法を用いて言語解析を行う言語解析処理システムにおいて、

解析対象の情報が付加されていないデータから、予め設定された問題表現の構造に合致する部分を抽出して問題表現相当部とする問題表現抽出処理手段と、

前記問題表現相当部を、問題と解とを含む教師データに変換する問題構造変換処理手段と、

前記教師データから素性と解との対を抽出し、抽出した素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データベースに保存する機械学習処理手段と、

入力されたデータから素性を抽出し、前記学習結果データベースに保存された前記学習結果をもとに解を推定する解推定処理手段とを備えることを特徴とする機械学習法を用いた言語解析処理システム。

【請求項２】　請求項１に記載の機械学習法を用いた言語解析処理システムにおいて、

前記機械学習処理手段は、多数の素性の重要度を各素性同士の従属性を考慮して自動で求める枠組みを用いて処理を行うことを特徴とする機械学習法を用いた言語解析処理システム。

【請求項３】　請求項１に記載の機械学習法を用いた言語解析処理システムにおいて、

前記機械学習処理手段は、前記教師データから素性と解の対を抽出して借用型教師信号とし、予め備えられた解析対象の情報が付加されたデータから素性と解との対を抽出して非借用型教師信号とし、前記借用型教師信号および前記非借用型教師信号を用いて機械学習を行うことを特徴とする機械学習法を用いた言語解析処理システム。

【請求項４】　機械学習法を用いた言語解析処理で用いる教師信号として借用する教師データを生成する教師データ生成処理方法において、

解析対象に関する情報が付加されていないデータから、予め設定された問題表現の構造に合致する部分を抽出して問題表現相当部とし、前記問題表現相当部を、問題と解とを含む教師データに変換する処理過程を備えることを特徴とする教師データ生成処理方法。

【請求項５】　機械学習法を用いて言語解析を行う言語解析処理方法において、

解析の問題と解とを含む教師データを記憶する教師データ記憶手段を備え、

前記教師データから素性と解との対を抽出し、抽出した素性と解との対を借用型教師信号として機械学習を行い、学習結果を学習結果データベースに保存する機械学習処理過程と、

入力されたデータから素性を抽出し、前記学習結果データベースに保存された学習結果をもとに解を推定する解推定処理過程とを備えることを特徴とする機械学習法を

2

用いた言語解析処理方法。

【請求項６】　請求項５に記載の機械学習法を用いた言語解析処理方法において、

前記機械学習処理過程は、多数の素性の重要度を各素性同士の従属性を考慮して自動で求める枠組みを用いて処理を行うことを特徴とする機械学習法を用いた言語解析処理方法。

【請求項７】　請求項５に記載の機械学習法を用いた言語解析処理方法において、

解析対象に関する解情報が付加されたデータを記憶する解データ記憶手段を備え、

前記機械学習処理過程は、前記教師データから素性と解の対を抽出して借用型教師信号とし、前記解情報を付加されたデータから素性と解との対を抽出して非借用型教師信号とし、前記借用型教師信号および前記非借用型教師信号を用いて機械学習を行うことを特徴とする機械学習法を用いた言語解析処理方法。

【請求項８】　機械学習法を用いて言い換えによる変形を含む省略解析を行う言語省略解析処理システムにおいて、

解析対象の情報が付加されていないデータから、予め設定された問題表現の構造に合致する部分を抽出して問題表現相当部とする問題表現抽出処理手段と、

前記問題表現相当部を、問題と解とを含む教師データに変換する問題構造変換処理手段と、

前記教師データから素性と解との対を抽出し、抽出した素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データベースに保存する機械学習処理手段と、

入力されたデータから素性を抽出し、前記学習結果データベースに保存された前記学習結果をもとに解を推定する解推定処理手段とを備えることを特徴とする機械学習法を用いた言語省略解析処理システム。

【請求項９】　請求項８に記載の機械学習法を用いた言語省略解析処理システムにおいて、

前記機械学習処理手段は、多数の素性の重要度を各素性同士の従属性を考慮して自動で求める枠組みを用いて処理を行うことを特徴とする機械学習法を用いた言語省略解析処理システム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、言語解析処理に関し、特に、機械学習法を用いた言語解析処理システム、教師データ生成処理方法、機械学習法を用いた言語解析処理方法、機械学習法を用いた言語省略解析処理システムに関する。

【０００２】

【従来の技術】言語解析処理の分野では、形態素解析、構文解析の次の段階である意味解析処理が重要性を増している。特に意味解析の主要部分である格解析処理、省

略解析処理などにおいて、処理にかかる労力の負担軽減や処理精度の向上が望まれている。

【０００３】格解析処理とは、文の一部が主題化もしくは連体化などをすることにより隠れている表層格を復元する処理を意味する。例えば、「りんごは食べた。」という文において、「りんごは」の部分は主題化しているが、この部分を表層格に戻すと「りんごを」である。このような場合に、「りんごは」の「は」の部分を「ヲ格」と解析する。

【０００４】また、「昨日買った本はもう読んだ。」という文において、「買った本」の部分が連体化しているが、この部分を表層格に戻すと「本を買った」である。このような場合にも、「買った本」の連体の部分を「ヲ格」と解析する。

【０００５】省略解析処理とは、文の一部に省略されている表層格を復元する処理を意味する。「みかんを買いました。そして食べました。」という文において、「そして食べました」の部分に省略されている名詞句（ゼロ代名詞）は「みかんを」であると解析する。

【０００６】本発明に関連する従来技術として、以下のような研究があった。

【０００７】格解析手法として、以下の参考文献１に示すような既存の格フレームを利用するものがある。

［参考文献１］Sadao Kurohashi and Makoto Nagao, A Method of Case Structure Analysis for Japanse Sentences based on Examples in Case Frame Dictionary, IEICE Transactions on Information and Systems, Vol.E77-D, No.2, (1994), pp227-239

また、以下の参考文献２に示すように、格解析において、解析対象としている分類や情報の付加を行っていないコーパス（以下、「生コーパス」という。）から格フレームを構築し、それを利用するものがある。

［参考文献２］河原大輔，黒橋禎夫，用言と直前の格要素の組を単位とする格フレームの自動獲得，情報処理学会，自然言語処理研究会，2000-NL-140-18，(2000)また、以下の参考文献３に示すように、格解析において、格情報付きコーパスを用いずに生コーパスでの頻度情報を利用して、最尤推定により格を求めるものがある。

［参考文献３］阿部川武，白井清昭，田中穂積，徳永健伸，統計情報を利用した日本語連体修飾語の解析，言語処理学会年次大会，(2001)，pp269-272なお、以下の参考文献４に示すように、格情報つきコーパスを用いた機械学習法としてｋ近傍法の一種のＴｉＭＢＬ法（参考文献５参照）を用いたものなどがある。

［参考文献４］Timothy Baldwin, Making lexical sense of japanese-englishmachine translation: A disambiguation extravaganza, Technical report,(Tokyo Institute of Technology, 2001), Technical Report, ISSN 0918-2802

［参考文献５］Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch, Timbl: Tilburg memory based learner version 3.0 reference guide, Technical report,(1995), ILK Technical Report-ILK 00-01

なお、参考文献３に示された阿部川らの研究や、参考文献４に示されたBaldwin の研究では、連体化の格解析処理のみを扱うものである。

【０００８】

【発明が解決しようとする課題】従来、日本語格解析を行う場合に用例とする格情報付きのコーパスに対し格情報を人手で付与していた。しかし、人手で解析規則や解析情報を付与することは、規則の拡張や規則の調節にかかる人的資源の問題や労力負担が大きいという問題がある。

【０００９】この点、教師付き機械学習法を言語解析処理に用いることは有効である。教師付き機械学習法では、解析対象となる情報が付与されたコーパスが教師信号として用いられている。しかし、この場合でも、コーパスに解析対象の情報を付加するという労力負担を軽減する必要がある。

【００１０】また、処理精度を向上させるために、なるべく多くの教師信号を使用できるようにすることが必要である。参考文献３の阿部川らの研究や、参考文献４のBaldwin の研究は、格情報のついていない生コーパスを用いて格解析処理を行うものである。ただし、これらの技術は連体化のみを扱う格解析処理である。

【００１１】機械学習法での教師信号を借用するため解析対象となる情報がついていない生コーパスなどを用いた機械学習法（以下、「教師信号借用型機械学習法」とよぶ。）を、より広範な言語処理において用いることができるようにすることが要求されている。

【００１２】そこで、格解析処理が省略解析処理と等価であることに着目し、省略解析処理において教師信号借用型機械学習法を用いた方法を提案する。

【００１３】また、動詞省略補完（参考文献６参照）、質問応答システム（参考文献７～９参照）などのより広範な言語解析について教師信号借用型機械学習法を用いた処理方法を提案する。

［参考文献６］村田真樹、長尾真、日本語文章における表層表現と用例を用いた動詞の省略の補完、言語処理学会誌、Vol.5, No.1,(1998)

［参考文献７］Masaki Murata, Masao Utiyama, and Hitoshi Isahara, Question answering system using syntactic information,(1999)

［参考文献８］村田真樹、内山将夫、井佐原均、類似度に基づく推論を用いた質問応答システム、自然言語処理研究会 2000-NL-135, (2000), pp181-188

［参考文献９］村田真樹、内山将夫、井佐原均、質問応答システムを用いた情報抽出、言語処理学会第６回年次大会ワークショップ論文集，(2000)，pp33-40

5

また、処理精度をより向上させるために、前記の教師信号借用型機械学習法により借用された教師信号と、解析対象である情報が付与されたデータを用いた教師あり機械学習法（以下、非借用型機械学習法という。）における教師信号とを併用した機械学習法（以下、併用型機械学習法という。）を用いた言語解析処理を提案する。

【００１４】また、省略解析の補完処理では語の生成を行うことから、前記の併用型機械学習法を用いた生成処理を提案する。

【００１５】本発明にかかる教師信号借用型機械学習法法もしくは併用型機械学習法は、教師あり機械学習法を用いている。本発明における教師あり機械学習法は、特に、各素性の重要度を、素性間の従属的関係を考慮した枠組みを用いて算出する過程を含むものである。この点、一般的に機械学習法として分類される方法のうち、各素性の類似度すなわち従属度を自ら決定しかかる算出過程を含まない場合のｋ近傍法、各素性の独立性を前提として素性間の従属性を考慮しないシンプルベイズ法などとも異なる。また、本発明における教師あり機械学習法は、阿部川らの方法（参考文献３参照）における、生コーパスで頻度による最尤推定とも異なる。最尤推定とは、固定文脈において頻度の最も大きいものを解とする手法であり、例えば格助詞を挟む体現と用言とを固定の文脈とする場合に、「りんご（？）食べる」の形をしているもので（？）の位置の助詞のうち最も頻度の高いものを解とするものである。

【００１６】以上のように、本発明の目的は、教師信号借用型機械学習法を用いた言い換えによる変形を含む言語省略解析処理システムを実現することである。

【００１７】さらに、好ましくは、前記教師信号借用型機械学習法として、各素性の重要度を素性間の従属的関係を考慮した枠組みを用いて算出する過程を含む機械学習法を用いた言語省略解析処理システムを実現することである。

【００１８】また、本発明の目的は、教師信号借用型機械学習法により借用した教師信号と、非借用型機械学習法の教師信号とによる機械学習法（併用型機械学習法）を用いた言語解析処理システムを実現することである。

【００１９】さらに、好ましくは、併用型機械学習法として、各素性の重要度を素性間の従属的関係を考慮した枠組みを用いて算出する過程を含む機械学習法を用いる言語解析処理システムを実現することである。

【００２０】本発明によれば、従来の教師信号以外に大量の教師信号を借用することができるため、使用する教師信号が増加し、よって学習の精度向上が期待できる。

【００２１】なお、本発明にかかる併用型機械学習法は、省略補完処理、文生成処理、機械翻訳処理、文字認識処理、音声認識処理など、語句を生成する処理を含むような極めて広範囲の問題に適用することができ、実用性の高い言語処理システムに用いることができる。

6

【００２２】

【課題を解決するための手段】上記の目的を達成するため、本発明は、機械学習法を用いて言語解析を行う言語解析処理システムにおいて、解析対象の情報が付加されていないデータから、予め設定された問題表現の構造に合致する部分を抽出して問題表現相当部とする問題表現抽出処理手段と、前記問題表現相当部を、問題と解とを含む教師データに変換する問題構造変換処理手段と、前記教師データから素性と解との対を抽出し、抽出した素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データベースに保存する機械学習処理手段と、入力されたデータから素性を抽出し、前記学習結果データベースに保存された前記学習結果をもとに解を推定する解推定処理手段とを備える。

【００２３】また、前記機械学習処理手段は、多数の素性の重要度を各素性同士の従属性を考慮して自動で求める枠組みを用いて処理を行う。

【００２４】また、前記機械学習処理手段は、前記教師データから素性と解の対を抽出して借用型教師信号とし、予め備えられた解析対象の情報が付加されたデータから素性と解との対を抽出して非借用型教師信号とし、前記借用型教師信号および前記非借用型教師信号を用いて機械学習を行う。

【００２５】また、本発明は、機械学習法を用いた言語解析処理で用いる教師信号として借用する教師データを生成する教師データ生成処理方法において、解析対象に関する情報が付加されていないデータから、予め設定された問題表現の構造に合致する部分を抽出して問題表現相当部とし、前記問題表現相当部を、問題と解とから構成される教師データに変換する処理過程を備える。

【００２６】また、本発明は、機械学習法を用いて言語解析を行う言語解析処理方法において、解析の問題と解とを含む教師データを記憶する教師データ記憶手段を備え、前記教師データから素性と解との対を抽出し、抽出した素性と解との対を借用型教師信号として機械学習を行い、学習結果を学習結果データベースに保存する機械学習処理過程と、入力されたデータから素性を抽出し、前記学習結果データベースに保存された学習結果をもとに解を推定する解推定処理過程とを備える。

【００２７】また、前記機械学習処理過程は、多数の素性の重要度を各素性同士の従属性を考慮して自動で求める枠組みを用いて処理を行う。

【００２８】また、本発明は、さらに、解析対象に関する解情報が付加されたデータを記憶する解データ記憶手段を備え、前記機械学習処理過程は、前記教師データから素性と解の対を抽出して借用型教師信号とし、前記解情報を付加されたデータから素性と解との対を抽出して非借用型教師信号とし、前記借用型教師信号および前記非借用型教師信号を用いて機械学習を行う。

【００２９】また、本発明は、機械学習法を用いて言い

換えによる変形を含む言語省略解析を行う言語省略解析処理システムにおいて、解析対象の情報が付加されていないデータから、予め設定された問題表現の構造に合致する部分を抽出して問題表現相当部とする問題表現抽出処理手段と、前記問題表現相当部を、問題と解とを含む教師データに変換する問題構造変換処理手段と、前記教師データから素性と解との対を抽出し、抽出した素性と解との対を教師信号として機械学習を行い、学習結果を学習結果データベースに保存する機械学習処理手段と、入力されたデータから素性を抽出し、前記学習結果データベースに保存された前記学習結果をもとに解を推定する解推定処理手段とを備える。

【００３０】また、前記機械学習処理手段は、多数の素性の重要度を各素性同士の従属性を考慮して自動で求める枠組みを用いて処理を行う。

【００３１】本発明は、解析対象用の教師信号のタグなどが付与されていないコーパスでも、問題が省略解析に類似する問題であるならば、その問題を教師信号として借用できることに着目し、この手法を単に格解析処理に用いるだけでなく、省略解析に類似するより広範な言語処理の問題においても利用できる手法を実現したものである。

【００３２】さらに、借用型でない本来の教師信号も併用する併用型機械学習法を提案して、処理負担の軽減と処理精度の向上とを図る処理方法を実現したものである。

【００３３】本発明の各処理手段または機能または要素は、コンピュータにインストールされ実行されるプログラムにより実現される。本発明を実現するプログラムは、コンピュータが読み取り可能な、可搬媒体メモリ、半導体メモリ、ハードディスクなどの適当な記録媒体に格納することができ、これらの記録媒体に記録して提供され、または、通信インタフェースを介して種々の通信網を利用した送受信により提供される。

【００３４】

【本発明の実施の形態】〔教師信号借用型機械学習法による処理〕教師信号借用型機械学習法を用いた言語解析処理のうち日本語解析処理を例に本発明の実施の形態を説明する。

【００３５】解析処理の一つである照応省略解析処理においては、照応省略に関する情報が付与されていないコーパスを利用することができると考える。その技術理論的背景を、以下の例を用いて示す。「例：みかんを買いました。これを食べました。」「用例ａ：ケーキを食べる。」、「用例ｂ：りんごを食べる。」このとき、「これ」の指示先を推定したいとする。この場合に、用例ａおよび用例ｂを使って、「を食べる」の前には食べ物についての名詞句がきそうであると予想し、この予想から「みかん」が指示先であると推定することができる。ここで、用例ａおよび用例ｂは、照応省略に関する情報が付与されていない普通の文でよい。

【００３６】一方、照応省略に関する情報が付与された用例を利用して解くことを考える。そのような用例は、例えば以下のような形をしている。「用例ｃ：りんごを買いました。これを食べました。（「これ」が「りんご」を指す。）」用例ｃでは、「りんごを買いました。これを食べました。」という文に対して、その文の「これ」が「りんご」を指すという照応省略に関する情報を付与しておくのである。このような用例ｃを用いることでも、「りんご」を指す例があるのなら、「みかん」も指すだろうと判断して、「みかん」を指示先を推定することができる。

【００３７】しかし、用例ｃのように、照応省略に関する情報をコーパスに付与することは大変労力のいることである。したがって、本発明のように、用例ｃの照応省略に関する情報を用いずに、照応省略に関する情報が付与されていない用例ａおよび用例ｂを用いることでも問題を解くことができるのならば、その方がコストが小さく、その意味で照応省略に関する情報が付与されていない用例を解析に利用できることは価値がある。

【００３８】このような解析対象に関する情報が付与されていない用例を用いた省略解析の例を以下に示す。

【００３９】（１）指示詞・代名詞・ゼロ代名詞照応解析

例：「みかんを買いました。そして {φを} 食べました。」

用例：「 {りんご} を食べる。」

指示詞・代名詞・ゼロ代名詞照応解析は、既に説明したように、指示詞や代名詞、文中で省略された代名詞（φ＝ゼロ代名詞）の指示先を推定するような解析である。以下の参考文献１０において詳細に説明している。［参考文献１０］村田真樹、長尾真、用例や表層表現を用いた日本語文章中の指示詞・代名詞・ゼロ代名詞の指示対象の推定、言語処理学会誌、Vol.4, No.1(1997)

（２）間接照応解析

例：「家がある。 {屋根} は白い。」

用例：「 {家} の屋根」

間接照応解析は、「ＡのＢ」の形をした用例を利用することで、「屋根」が前文の「家」の屋根であると推定するような解析である。以下の参考文献１１において詳細に説明している。

［参考文献１１］村田真樹、長尾真、意味的制約を用いた日本語名詞における間接照応解析、言語処理学会誌、Vol.4, No.2, (1997)

（３）動詞の省略補完

例：「そううまくいくとは」

用例：「そんなにうまくいくとは {思えない} 。」

例「そううまくいくとは」の後ろに省略されている動詞部分を「そううまくいくとは」を含む文を集めて、その用例文を用いて推測するような解析である。前述の参考

文献６で説明している。

【００４０】（４）「ＡのＢ」の意味解析

例：「写真の人物」⇒「写真に描かれた人物」

用例：「写真に人物が描かれる」

「ＡのＢ」のような語句の意味関係は多様である。しかし、意味関係の中には動詞で表現できるものがある。そのような動詞は、名詞Ａ、名詞Ｂおよび動詞との共起情報から推測できる。「ＡのＢ」の意味解析とは、このような共起情報により意味関係を推測するような解析である。解析の詳細は、以下の参考文献１２に説明されている。

［参考文献１２］田中省作、富浦洋一、日高達、統計的手法を用いた名詞句「ＮＰのＮＰ」の意味関係の抽出、言語理解とコミュニケーション研究会　NLC98-4,(1998), pp23-30

（５）換喩解析

例：「漱石を読む。」⇒「漱石の小説を読む。」

用例：「漱石の小説」「小説を読む」

「漱石を読む」の「漱石」は「漱石が書いた小説」を意味する。換喩解析は、そのような省略された情報を、「ＡのＢ」「ＣをＶする」という形をした用例を組み合わせて用いることで補完する解析である。以下の参考文献１３および参考文献１４において説明している。

［参考文献１３］村田真樹、山本専、黒橋禎夫、井佐原均、長尾真、名詞句「ａのｂ」「ａｂ」を利用した換喩解析、実行知能学会誌、Vol.15, No.3 (2000)

［参考文献１４］内山将夫、村田真樹、馬青、内元清貴、井佐原均、統計的手法による換喩の解釈、言語処理学会誌, Vol.7, No.2, (2000)

（６）連体化した節の格解析

例：「オープンする施設」⇒格関係＝ガ格

用例：「施設がオープンする」

連体化した節の格解析とは、名詞と動詞の共起情報を用いて隠れている連体化した節の格を推定する解析である。解析の内容は前記の参考文献３に詳しく説明されている。

【００４１】図１に、本発明にかかるシステムの構成例を示す。図１中、１は本発明にかかる言語解析処理システムを表す。言語解析処理システム１は、ＣＰＵ、メモリなどで構成され、問題表現相当部抽出部１１、問題表現情報記憶部１２、問題構造変換部１３、意味解析情報記憶部１４、教師データ記憶部１５、解－素性対抽出部１７、機械学習部１８、学習結果データベース１９、素性抽出部２１、解推定処理部２２を持つ。

【００４２】問題表現相当部抽出部１１は、予め、どのようなものが問題表現に相当する部分であるかを記憶した問題表現情報記憶部１２を参照して、解析対象の情報が付与されていない生コーパス２から入力された文について、問題表現に相当する部分を抽出する手段である。

【００４３】問題表現情報記憶部１２は、前記（１）～（６）に示すような省略解析の問題表現を予め記憶しておく。また、意味解析の場合に用いる意味解析情報は、予め意味解析情報記憶部１４に記憶しておく。

【００４４】問題構造変換部１３は、問題表現相当部抽出部１１で抽出された入力文の問題表現に相当する部分を解として抽出し、さらに、その部分を問題表現に変換し、変換結果の文を問題とし、かつ、抽出した解を解とする教師データを教師データ記憶部１５に記憶する手段である。

【００４５】また、問題構造変換部１３は、問題表現に変換した結果である文を変形する必要がある場合に、意味解析情報記憶部１４を参照して、当該結果文を変形したものを問題とする。

【００４６】解－素性対抽出部１７は、問題－解の構造を持つ教師データを記憶する教師データ記憶部１５から、事例ごとに、事例の解と素性の集合との組を抽出する手段である。

【００４７】機械学習部１８は、解－素性対抽出部１７により抽出された解と素性の集合の組から、どのような素性のときにどのような解になりやすいかを機械学習法により学習し、その学習結果を学習結果データベース１９に保存する手段である。

【００４８】素性抽出部２１は、入力されたデータ３から、素性の集合を抽出し、解推定処理部２２へ渡す手段である。

【００４９】解推定処理部２２は、学習結果データベース１９を参照して、素性抽出部２１から渡された素性の集合の場合に、どのような解になりやすいかを推定し、推定結果である解析情報４を出力する手段である。

【００５０】以下に、本発明の処理の流れを説明する。

【００５１】図２に、教師データの生成処理の処理フローチャートを示す。

【００５２】ステップＳ１：まず、生コーパス２から解析対象の情報がなにも付与されていない普通の文が問題表現相当部抽出部１１に入力される。

【００５３】ステップＳ２：問題表現相当部抽出部１１では、生コーパス２から入力された普通文の構造を検出し、入力された普通文から問題表現に相当する部分を抽出する。このとき、どのようなものが問題表現相当部であるかの情報は、問題表現情報記憶部１２に記憶されている問題表現情報により与える。すなわち問題表現の構造と検出した普通文の構造とのマッチングを行い、一致するものを問題表現相当部とする。

【００５４】ステップＳ３：問題構造変換部１３では、問題表現相当部抽出部１１で抽出された問題表現相当部を解として抽出し、その部分を問題表現に変換する。そして、変換結果の文を問題とし抽出した解を解とする教師データを教師データ記憶部１５に記憶する。

【００５５】なお、問題構造変換部１３では、問題表現に変換する際に、意味解析情報を必要とする場合には、

予め意味解析情報記憶部１４に記憶されている意味解析情報を参照する。

【００５６】具体的には、以下のような処理を行う。

【００５７】例えば、前述（３）に示す動詞の省略補完の場合には、問題表現情報記憶部１２には、文末の動詞部分が問題表現相当部として記述されている。そして、生コーパス２から、「そんなにうまくいくとは思えない」という文が入力されると、問題表現相当部抽出部１１では、文末の動詞「思えない」が問題表現相当部であると認識する。

【００５８】問題構造変換部１３では、文末の動詞「思えない」を解として抽出し、元の文の動詞「思えない」の部分を”省略された動詞”という記号に置き換える。この結果、「問題⇒解」：「そんなにうまくいくとは”省略された動詞”」⇒「思えない」という教師データが得られるので、この教師データを教師データ記憶部１５へ記憶する。

【００５９】そして、この教師データは、文脈：「そんなにうまくいくとは」，分類先：「思えない」という形式の機械学習法で用いる教師信号とすることができる。すなわち、解－素性対抽出部１７では、教師データを文脈から分類先を学習する教師あり機械学習の問題として使用することができる。

【００６０】また、前述（１）の格解析の場合には、問題表現情報記憶部１２には、格助詞が問題表現相当部として記述されている。そして、生コーパス２から、「りんごを食べる」という文が入力されると、問題表現相当部抽出部１１では、格助詞「を」が問題表現相当部として認識する。

【００６１】問題構造変換部１３では、格助詞「を」を解として抽出し、元の文の格助詞「を」の部分を”認識すべき格”という記号に置き換える。この結果、「問題⇒解」：「りんご”認識すべき格”食べる」⇒「を」という教師データが得られるので、この教師データを教師データ記憶部１５へ記憶する。この場合も同様に、解－素性対抽出部１７を介して、文脈：「食べる」，分類先：「りんごを」という教師信号となる。

【００６２】前述した他の解析例についても、同様の処理を行い、それぞれの教師データを出力する。そして、例えば、前述（２）の間接照応解析の場合には、文脈：「の屋根」，分類先：「家」という教師信号に、また、前述（４）の「ＡのＢ」の意味解析の場合には、文脈：「写真」「人物」，分類先：「描かれる」という教師信号に、また、前述（５）の換喩解析の場合には、文脈：「漱石の」，分類先：「小説」文脈：「を読む」，分類先：「小説」という教師信号に、また、前述（６）の連体化における格解析の場合は、文脈：「施設」「オープンする」，分類先：「ガ格」という教師信号になる。

【００６３】このように、省略解析と解釈できる問題表現については、解析対象用のタグがついていない生コーパス２を機械学習方法の教師信号とすることができる。

【００６４】特に、単純な省略補完だけではなく、例えば「オープンする施設」を「施設がオープンする」ととらえる格解析のように、言葉を少し補って言い換えて解釈するような問題についても、生コーパス２を機械学習方法の教師信号とすることができる。すなわち、意味解釈の問題は、たいていの場合、言い換えた文によってその答えを表現するため、本発明は言葉を少し補いながら言い換えて解釈するような問題一般も適用範囲に含めることができることを意味する。一例として、本発明を質問応答システムに適用する場合について説明する。

【００６５】質問応答システムでの質問応答は、疑問詞の部分が省略しておりこの部分を補完する問題であると考えることができる。この場合に、よく似た文を集めてその文の疑問詞にあたる部分を解答として出力する（参考文献７～９参照）。

【００６６】例えば、以下のような質問および解答の事例の場合に、事例：「日本の首都はどこですか」⇒解答＝東京用例：「日本の首都は東京です」という教師データは、文脈：「日本の首都は」，分類先：「東京」文脈：「の首都は東京です」，分類先：「日本」という教師信号になる。

【００６７】このように、教師データ記憶部１５に記憶される教師データは、通常の教師信号の形式と同じような構造になっているため、教師あり機械学習法の教師信号として用いることができ、さまざまな高度な手法が提案されている機械学習法の中から最適な手法を選択して問題を解くことができる。

【００６８】また、機械学習法では、解析に用いる情報をかなり自由に定義することができることから、広範な情報を教師信号として利用でき、結果的に解析精度が向上しやすい。

【００６９】図３に、教師データを教師信号とする機械学習法による解析処理の処理フローチャートを示す。

【００７０】ステップＳ１１：まず、解－素性対抽出部１７では、教師データ記憶部１５から、各事例ごとに、解と素性の集合との組を抽出する。素性とは、解析に用いる情報の細かい１単位を意味する。解－素性対抽出部１７は、素性の集合を機械学習に用いる文脈とし、解を分類先とする。

【００７１】ステップＳ１２：続いて、機械学習部１８では、抽出された解と素性の集合との組から、どのような素性のときにどのような解になりやすいかを機械学習し，その学習結果を学習結果データベース１９に保存する。

【００７２】機械学習の手法は、多数の素性の重要度を各素性同士の従属性を考慮して自動で求める枠組みを用いて算出する処理過程を含むものであればよい。例えば、以下に示すような決定リスト法、最大エントロピー法、サポートベクトルマシン法などを用いるが、これら

の手法に限定されない。

【００７３】決定リスト法は、素性（解析に用いる情報で文脈を構成する各要素）と分類先の組を規則とし、それらをあらかじめ定めた優先順序でリストに蓄えおき、解析すべき入力が与えられたときに、リストで優先順位の高いところから入力のデータと規則の素性を比較し素性が一致した規則の分類先をその入力の分類先とする方法である。

【００７４】最大エントロピー法は、あらかじめ設定しておいた素性ｆ$_j$（１≦ｊ≦ｋ）の集合をＦとするとき、所定の条件式を満足しながらエントロピーを意味する式を最大にするときの確率分布ｐ（ａ，ｂ）を求め、その確率分布にしたがって求まる各分類の確率のうち、もっとも大きい確率値を持つ分類を求める分類とする方法である。

【００７５】サポートベクトルマシン法は、空間を超平面で分割することにより、２つの分類からなるデータを分類する手法である。

【００７６】本形態では、最も処理精度の高いサポートベクトルマシン法を用いた処理例についての詳細な説明を後述する。

【００７７】決定リスト法および最大エントロピー法については、以下の参考文献１５に説明している。［参考文献１５］村田真樹、内山将夫、内元清貴、馬青、井佐原均、種々の機械学習法を用いた多義解消実験、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，NCL2001-2，(2001)ステップＳ１３：解を求めたいデータ３が素性抽出部２１に入力される。

【００７８】ステップＳ１４：素性抽出部２１では，解－素性対抽出部１７での処理とほぼ同様に，入力されたデータ３から素性の集合を取り出し，それらを解推定処理部２２へ渡す。

【００７９】ステップＳ１５：解推定処理部２２では，渡された素性の集合の場合にどのような解になりやすいかを学習結果データベース１９をもとに特定し、特定した解である解析情報４を出力する。

【００８０】例えば，データ３が「りんごは食べる」であって、解析したい問題が「認識すべき格」であれば、「ヲ格」という格情報を出力する。また、データ３が「そんなにうまくいくとは」であって、解析したい問題が「補完すべき動詞」であれば、省略された動詞「思えない」を出力する。

【００８１】図４に、機械学習法としてサポートベクトルマシン法を用いる場合のシステム構成例を示す。図４に示す言語解析処理システム５の構成例は、図１に示す構成例とほぼ同様である。図４において、図１に示す手段と同一の機能を持つ手段には同一の番号を付与している。

【００８２】素性－解対・素性－解候補対抽出部５１は、教師データ記憶部１５から、事例ごとに、事例の解もしくは解候補と事例の素性の集合との組を抽出する手段である。ここで、解候補とは、解以外の解の候補を意味する。

【００８３】機械学習部５２は、素性－解対・素性－解候補対抽出部５１により抽出された解もしくは解候補と素性の集合との組から、どのような解もしくは解候補と素性の集合のときに、正例である確率または負例である確率を、例えばサポートベクトルマシン法により学習し、その学習結果を学習結果データベース５３に保存する手段である。

【００８４】素性－解候補抽出部５４は、入力されたデータ３から、解候補と素性の集合とを抽出し、解推定処理部５５へ渡す手段である。

【００８５】解推定処理部５５は、学習結果データベース５３を参照して、素性－解候補抽出部５４から渡された解候補と素性の集合との場合に、正例または負例である確率を求めて、正例である確率が最も大きい解候補を解析情報４として出力する手段である。

【００８６】サポートベクトルマシン法を説明するため、図５に、サポートベクトルマシン法のマージン最大化の概念を示す。図５において、白丸は正例、黒丸は負例を意味し、実線は空間を分割する超平面を意味し、破線はマージン領域の境界を表す面を意味する。図５（Ａ）は、正例と負例の間隔が狭い場合（スモールマージン）の概念図、図５（Ｂ）は、正例と負例の間隔が広い場合（ラージマージン）の概念図である。

【００８７】このとき、2 つの分類が正例と負例からなるものとすると、学習データにおける正例と負例の間隔（マージン）が大きいものほどオープンデータで誤った分類をする可能性が低いと考えられ、図５（Ｂ）に示すように、このマージンを最大にする超平面を求めそれを用いて分類を行なう。

【００８８】サポートベクトルマシン法は基本的には上記のとおりであるが、通常、学習データにおいてマージンの内部領域に少数の事例が含まれてもよいとする手法の拡張や、超平面の線形の部分を非線型にする拡張（カーネル関数の導入）がなされたものが用いられる。

【００８９】この拡張された方法は、以下の識別関数を用いて分類することと等価であり、その識別関数の出力値が正か負かによって二つの分類を判別することができる。

【００９０】

【数１】

15　　16

$$f(\mathbf{x}) = sgn\left(\sum_{i=1}^{l}\alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b\right) \quad (1)$$

$$b = -\frac{max_{i,y_i=-1}b_i + min_{i,y_i=1}b_i}{2}$$

$$b_i = \sum_{j=1}^{l}\alpha_j y_j K(\mathbf{x}_j,\mathbf{x}_i)$$

【００９１】ただし、ｘは識別したい事例の文脈（素性の集合）を、x_i とy_j （ｉ＝１，…，ｌ，y_j ∈｛１，－１｝）は学習データの文脈と分類先を意味し、関数ｓｇｎは、

ｓｇｎ（ｘ）＝１（ｘ≧０）
　　－１（otherwise ）　　（２）

であり、また、各α_i は式（４）と式（５）の制約のもと式（３）を最大にする場合のものである。

【００９２】

【数２】

$$L(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j K(\mathbf{x_i},\mathbf{x_j}) \quad (3)$$

$$0 \le \alpha_i \le C \ (i = 1,...,l) \quad (4)$$

$$\sum_{i=1}^{l}\alpha_i y_i = 0 \quad (5)$$

【００９３】また、関数Ｋはカーネル関数と呼ばれ、様々なものが用いられるが、本形態では以下の多項式のもの

Ｋ（ｘ，ｙ）＝（ｘ・ｙ＋１）^{d}　　（６）

を用いる。

Ｃ、ｄは実験的に設定される定数である。後述する具体例ではＣはすべての処理を通して１に固定した。また、ｄは、１と２の二種類を試している。ここで、$\alpha_i > 0$ となるx_i は，サポートベクトルと呼ばれ、通常，式（１）の和をとっている部分はこの事例のみを用いて計算される。つまり、実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない。

【００９４】

【００９５】なお、拡張されたサポートベクトルマシン法の詳細については、以下の参考文献１６および参考文献１７を参照されたい。

［参考文献１６］Nello Cristianini and John Shawe-Taylor, An Introductionto Support Vector Machines and other kernel-based learning methods,(Cambridge University Press,2000)

［参考文献１７］Taku Kudoh, Tinysvm:Support Vector machines,(http://cl.aist-nara.ac.jp/taku-ku//software/Tiny SVM/index.html,2000)

サポートベクトルマシン法は、分類の数が２個のデータを扱うもので、通常これにペアワイズ手法を組み合わせて用いることで、分類の数が３個以上のデータを扱うことになる。

【００９６】ペアワイズ手法とは、Ｎ個の分類を持つデータの場合に、異なる二つの分類先のあらゆるペア（Ｎ（Ｎ－１）／２個）を作り、各ペアごとにどちらがよいかを２値分類器（ここではサポートベクトルマシン法によるもの）で求め、最終的にＮ（Ｎ－１）／２個の２値分類器の分類先の多数決により、分類先を求める方法である。

【００９７】本形態における２値分類器としてのサポートベクトルマシンは、サポートベクトルマシン法とペアワイズ手法を組み合わせることによって実現するものであり、以下の参考文献１８により工藤氏が作成したＴｉｎｙＳＶＭを利用している。

［参考文献１８］工藤拓　松本裕治, Support vector machineを用いたchunk 同定、自然言語処理研究会、 2000-NL-140,(2000)

図６に、機械学習法としてサポートベクトルマシン法を用いた解析処理の処理フローチャートを示す。

【００９８】ステップＳ２１：素性－解対・素性－解候補対抽出部５１では、各事例ごとに、解もしくは解候補と素性の集合との組を抽出する。解と素性の集合との組を正例、解候補と素性の集合との組を負例とする。

【００９９】ステップＳ２２：機械学習部５２では、解もしくは解候補と素性の集合との組から、どのような解もしくは解候補と素性の集合のときに正例である確率または負例である確率であるかを機械学習法例えばサポートベクトルマシン法により学習する。そして、その学習結果を学習結果データベース５３に保存する。

【０１００】ステップＳ２３：素性－解候補抽出部５４に、解を求めたいデータ３が入力される。

【０１０１】ステップＳ２４：素性－解候補抽出部５４では、入力されたデータ３から解候補と素性の集合との組を取り出し、解推定処理部５５へ渡す。

【０１０２】ステップＳ２５：解推定処理部５５では、渡された解候補と素性の対の場合に、正例である確率および負例である確率を求める。この確率は、すべての解候補について計算する。

【０１０３】ステップＳ２６：解推定処理部５５では、すべての解候補の中から、正例である確率が最も大きい解候補を求め、その解候補を解とする解析情報４を出力する。

【０１０４】〔借用型教師信号を用いた機械学習法による処理〕教師データ記憶部１５に記憶される教師データは、通常の教師信号の形式である「問題⇨解」となっている。このため、本来の解析対象用のタグのついたコーパスからデータをとった教師信号（非借用型教師信号）と同時に併用して用いることができる。教師データと、非借用型教師信号とを併用すれば、多くの情報を利用できるために機械学習の結果の精度が向上する。

【０１０５】ところで、照応解析などでは、指示先が本文にあり用例だけの情報で指示先を特定するのは困難な問題もあるため、借用した教師データだけを用いて解析を行なうことはできない場合もある。このような場合には、非借用型教師信号、すなわち従来の教師信号も用いる併用型機械学習法を用いた処理によって対処することができる。

【０１０６】用例「りんごも食べる」について、生成する教師データとして、「問題⇨解」：「リンゴ”認識すべき格”食べる」⇨「を」が得られる。一方、本来の教師信号では、「問題⇨解」：「りんごも食べる」⇨「を」であることを考えると、「も」と”認識すべき格”の部分が少し異なる。「も」も”認識すべき格”の一部ではあるが、本来の教師信号の「も」の方が、「も」があるだけ情報量が多いことになる。すなわち、非借用型教師信号の方が情報が多い。よって、併用型教師あり機械学習法による処理がよりよいと考えられる。

【０１０７】また、格解析でも、つねに表層格を補完するのではなく、表層格を用いた文に変形できないことから、外の関係（格関係にならない関係）などは教師データでは扱えない問題もある。

【０１０８】しかし、ここで格解析というしばりを排して言い換えによる文解釈という立場から見れば、外の関係も教師データを用いた機械学習で扱えることになる。例えば、外の関係の文「さんまを焼くけむり」は、「さんまを焼く時に出るけむり」と言い換えて解釈できる場合がある。「さんまを焼く時に出るけむり」と言い換える解釈を正解とする問題設定であるならば、連体節とその係り先の名詞との間の省略された表現「時に出る」を補完するという省略補完の問題となり、借用型教師データを用いた機械学習で扱える問題となり、併用型機械学習法による処理に適している。

【０１０９】また、省略解析だけでなく、生成についても取り扱うことができると考える。教師信号借用型機械学習法すなわち、解析対象とするタグがふられていないコーパスを用いることができるという点で、省略解析と生成とが似ていることについては、以下の参考文献１９で指摘した。

［参考文献１９］村田真樹、長尾真、表層表現と用例を用いた照応省略解析手法、言語理解とコミュニケーション研究会　NCL97-56,(1997)

例えば格助詞の生成の例を示す。格助詞の生成では、例えば問題－解の組は、「問題⇨解」：「りんご←（obj）－食べる」⇨「を」といったものになる。生成の場合に、一般に生成される部分の意味を深層格など（例：obj）を用いて表現する。ここで、obj とは目的格を意味する。この問題－解の組は、このobj の部分が格助詞の生成の結果では「を」になるということを示しており、前述でいう非借用型教師信号に相当する。

【０１１０】また、この問題での借用型教師信号は、解析対象とするタグがふられていない生コーパス２から「りんごを食べる」といった文を取り出して、それを借用型教師信号として扱うことで以下のようなものとなる。

【０１１１】「問題⇨解」：「りんご”生成すべき格”食べる」⇨「を」これらの非借用型教師信号と借用型教師信号とは非常に類似しており、「obj」と”生成すべき格”の部分とが少し異なるだけで借用型教師信号も非借用型教師信号と同様に教師信号として十分に用いることができる。つまり、格助詞の生成においても教師信号借用型機械学習法を用いることができる。

【０１１２】また「obj 」と”生成すべき格”の部分とでは、「obj 」の方が、「obj 」があるだけ情報量が多い。このため、この問題においても、本来の教師信号、すなわち非借用型教師信号の方が情報が多いことになる。したがって、借用型教師信号だけでなく非借用型教師信号を用いる併用型機械学習法による処理を用いる方がよりよい。

【０１１３】また、英日機械翻訳における格助詞生成の例を示す。この問題では、問題－解の組は、「問題⇨解」：「eat → apple」⇨「を」のように与えられる。これは、 "I eat apple." という文の eatと appleの関係が、英語から日本語に変換すると「を」になるということを示しており、非借用型教師信号に相当するものである。この問題でも解析対象とするタグがふられていない生コーパス２から「りんごを食べる」といった文を取り出して、それを借用型教師信号として扱うことで、「問題⇨解」：「りんご”生成すべき格”食べる」⇨「を」となる。

19

【０１１４】ここで、問題をみると、本来の教師信号（非借用型教師信号）と借用型教師信号とは、全然一致する部分がないことがわかる。このままでは借用型教師信号は役に立たない。そこで、それぞれの信号について問題部分は英日もしくは日英翻訳しておく。そうすると「問題⇨解」：「 eat （食べる）→ apple（りんご）」⇨「を」「問題⇨解」：「りんご（ apple)”生成すべき格”食べる（eat)」⇨「を」のようになる。この状態であれば少々は一致するため、借用型教師信号も教師信号として役に立つ。例えば、単語を切り出して、それらを学習に用いる素性とする場合に、それらは「eat 」、「apple 」、「食べる」、「りんご」であり、ほとんど一致する。

【０１１５】また、機械翻訳では各部分の翻訳の候補を組み合わせて全体の翻訳を組み合わせることもあり、他の部分の翻訳を先に処理することを前提にすれば「eat →apple 」の部分を「食べる→りんご」などにすでになっていることを前提として「問題⇨解」：「食べる→りんご」⇨「を」という教師信号になっていると扱ってもよい。

【０１１６】この場合も本来の教師信号の問題部分と借用する教師信号とに一致部分があるため、併用型機械学習法を利用することができる。

【０１１７】また、各部分の翻訳の候補を組み合わせて全体の翻訳を組み合わせる際に、各部分の翻訳の候補を複数残しておいて、それらの組み合わせの分をすべて解候補として残しながら解を求めていくようにしてもよい。このように翻訳の候補を解候補として扱うようにしても、上記のように自分（この場合「を」）以外の部分（この場合は、「食べる」および「りんご」）の翻訳結果を利用することができる。

【０１１８】併用型機械学習法による処理の場合に、図１または図４に示すシステム構成例において、解データベース１６を予め用意しておく必要がある。解データベース１６は、従来の教師あり機械学習法で用いられる、解析情報を人手などにより付与したコーパスなどである。そして、図１に示すシステムの場合に、解－素性対抽出部１７は、教師データ記憶部１５および解データベース１６から、各事例ごとに解と素性の集合との組を抽出する。また、図４に示すシステムにおいても、素性－解対・素性－解候補対抽出部５１は、同様に、教師データ記憶部１５および解データベース１６から、各事例ごとに解もしくは解候補と素性の集合との組を抽出する。

【０１１９】〔具体例〕本形態における具体的な処理例について説明する。

【０１２０】具体例での格解析の問題設定と素性（解析に用いる情報）について、すなわち機械学習に用いる文脈（素性の集合）と分類先を説明する。格解析を行なう対象は以下のものとした。・連体化した節の用言とその係り先の体言との間の関係・格助詞のみがつく体言、助詞が一切つかない体言を除く体言が用言にかかる場合のその体言と用言との関係（例えば、「この問題｛さえ｝解かれた。」)また、分類先として、ガ格、ヲ格、ニ格、デ格、ト格、カラ格（６分類）およびその他（外の関係、格関係にならない主題など）の７つの分類を用いた。このとき、受け身の文の場合でも受け身の文型のまま表層格の推定を行なうこととした。例えば「解かれた問題」の場合には、「問題が解かれた」となるのでガ格として扱う。受け身を能動態に直して「問題を解く」と解釈してヲ格とはしなかった。

20

【０１２１】また、外の関係とは、関係節の用言と係り先の体言が格関係にならない場合のことをいう。例えば、「さんまを焼くにおい」の文の「焼く」と「におい」とは格関係が成立しないので、このような文は外の関係と呼ばれる。

【０１２２】また、連体化以外で「その他」の分類とするものに、例えば、「｛九一年も｝出生数が前年より千六百六十人多かった」の「九一年も」がある。この「九一年も」は、ガガ文としてガ格としてもよい場合もあるからである。

【０１２３】また、以下の「過去一年間に｛三度も｝首相が代わる」の「三度も」のような副詞も「その他」の分類とした。

【０１２４】本例では、助詞「も」がなければ解析の対象としないこととした。助詞の脱落現象の少ない分野のデータならば、助詞が一つもついていなければ副詞と判断してもよいだろうが、助詞の省略が存在するとなると、助詞のついていない体言も係り先の用言と格関係を持つ可能性があるために、それらの体言もすべて解析対象とする必要があるためである。

【０１２５】また、文脈としては以下のものを定義した。ただし、体言ｎと用言ｖの間の格関係を求める場合として表している。

１．問題が連体節か主題化のものか主題化の場合は体言ｎについている助詞
２．用言ｖの品詞
３．用言ｖの単語の基本形
４．用言ｖの単語の分類語彙表の分類番号の1 、2 、3 、4 、5 、7 桁までの数字。ただし、分類番号に対して文献の表の変更を行なっている。
５．用言ｖにつく助動詞列（例：「れる」、「させる」）
６．体言ｎの単語
７．体言ｎの単語の分類語彙表の分類番号の1 、2 、3 、4 、5 、7 桁までの数字。ただし、分類番号に対して文献の表の変更を行なっている。
８．用言ｖにかかる体言ｎ以外の体言の単語列
ただし、どういった格でかかっているかの情報をAND でつけることとした。
９．用言ｖにかかる体言ｎ以外の体言の単語集合の分類

語彙表の分類番号の1 、2 、3 、4 、5 、7 桁までの数字。ただし、分類番号に対して文献の表の変更を行なっている。また、どういった格でかかっているかの情報をAND でつけることとした。

１０．用言ｖにかかる体言ｎ以外の体言がとっている格

１１．同一文に共起する語

本例では、以上の素性のいくつかを用いて行った。なお、教師信号借用型機械学習法を用いる場合は、前記１．の素性は用いることができない。

【０１２６】まず従来の教師あり機械学手法（非借用型機械学習法）を用いた処理を行なった。データは京都大コーパス中の毎日新聞９５年１月１日の一日分を用いた（参考文献２０参照）。［参考文献２０］黒橋禎夫、長尾真、京都大学テキストコーパス・プロジェクト、言語処理学会第３回年次大会、1997、pp115-118このデータに対し、前記したように定義した問題設定で分類先を付与した。京大コーパスの構文タグが誤っていると判明した部分はデータから除いた。事例数は１，５３０個であった。図７に、全事例における分類先の出現の分布を示す。この事例の分布から、コーパスの用例中、ガ格が圧倒的に多く、ついで連体における外の関係が多いことがわかる。

【０１２７】次に、教師信号借用型機械学習法を用いた処理を行なった。借用する教師データ用の用例は京大コーパス中の毎日新聞９５年１月１～１７日の１６日分（約２万文）を用いた。このデータのうち、体言と用言を係り受け関係を格助詞のみで結んでいるもののみを教師データとした。全事例数は５７，８５３個であった。このとき、前記の定義の素性のうち１．の素性は、主題化・連体化していないものからデータをもってくるために用いることができない。

【０１２８】機械学習法としては、ＴｉＭＢＬ法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法を用いた。ＴｉＭＢＬ法、シンプルベイズ法については、処理精度の比較のために用いた。

【０１２９】なお、ＴｉＭＢＬ法は、Daelemans らが開発したシステムで、類似するｋ個の事例でもとめるｋ近傍法を用いるものになっている（参考文献５参照）。さらにＴｉＭＢＬ法では事例間の類似度はあらかじめ定義しておく必要はなく、素性を要素とした重みつきのベクトルの間の類似度という形で自動的に算出される。また本稿ではｋ＝３を用いその他はデフォルトの設定で利用した。シンプルベイズ法は、あらかじめ類似度の定義を与えるｋ近傍法の一手法である。

【０１３０】まず、教師信号借用型機械学習法の基本性能を調べるために、表層格の再推定という問題を解く。これは文中の表層格を消して、それをもう一度推定できるか否かを試すものである。この問題を対象として、さきほどの借用型教師信号（５７，８５３個）で記事ごとの１０分割のクロスバリデーションを用いて実験した。

【０１３１】図８に、各手法の処理の結果（精度）を示す。ＴｉＭＢＬ、ＳＢ、ＤＬ、ＭＥ、ＳＶＭは、それぞれＴｉＭＢＬ法、シンプルベイズ法、決定リスト法、最大エントロピー法、サポートベクトルマシン法を意味する。図８に示すように、サポートベクトルマシン法（ＳＶＭ）がもっとも精度が良く、７割の精度を得た。

【０１３２】この処理の結果からも、文生成における助詞の生成については、少なくともこの精度で処理を行えることを示している。また、文生成の処理の場合には、併用型機械学習法を用いた処理を用いることにより、深層格などなんらかの格に対する情報を入力としても与えることができるため、図８に示す処理結果よりも高い精度が得られると考えられる。また、一般的な助詞脱落の補完問題は、この程度の処理精度を得ることができれば、解けるであろうことがわかる。

【０１３３】さらに、教師信号借用型機械学習法を用いて、最初に用意した主題化・連体化したデータで、表層格復元の処理を行なった。この場合には、借用型教師信号では外の関係などの「その他」の分類を推定することができないので、「その他」の分類の事例を除いて処理を行なった。そのため、評価用のデータの事例数は１，５３０から１，１８８に減少した。機械学習にはさきほど集めた借用型教師信号（５７，８５３個）を用いた。図９に、この処理の結果を示す。

【０１３４】また、この処理では、ガ格、ヲ格、ニ格、デ格の４つの格のそれぞれの精度の平均でも評価した。図１０に、この処理の結果を示す。

【０１３５】ここでは比較のために、この１，１８８事例を学習に用いた非借用型機械学習法による結果も示す。また、この１，１８８個の非借用教師信号と、５７，８５３個の借用教師信号の両方を併用する併用型機械学習法による結果も示す。ただし、これらの処理では記事を単位とする１０分割のクロスバリデーションを行ない、解析対象の事例と同じ記事の借用教師信号と非借用教師信号は用いないようにした。

【０１３６】結果より以下のことがわかる。まず、図９に示す処理結果の全事例での精度で検討する。機械学習法としてはサポートベクトルマシン法が一般的に最も良い。したがって、以降の検討ではサポートベクトルマシン法の結果のみを使うこととした。

【０１３７】借用型機械学習法での精度は５５．３９％であった。主な格の出現がガ格、ヲ格、ニ格、デ格の４つであったので、ランダムな選択の場合の処理精度は２５％であるから、これよりはよい結果となっている。借用した教師信号を用いた場合の精度としてはよいものと思われる。

【０１３８】併用型、借用型、非借用型の中では非借用型機械学習法が最もよかった。借用型教師信号としたデータは、実際の問題とは異なる性質を持っている可能性

がある。したがって、このようなデータを借用することにより、処理精度が低下する可能性は十分ありうる。図９に示す処理結果は、このような状況を反映したものと考えられる。

【０１３９】この処理の評価に用いたデータは１，１８８事例であり、そのうちガ格は１，０２５事例であり、ガ格の出現確率は８６．２８％である。したがって、何も考えずに、すべてガ格であると判定したと場合でも、８６．２８％の精度を得る。しかし、このような判定では、他の格の解析精度は０％であり、この処理結果は利用先によっては何も役に立たない可能性がある。そこで、図１０に示す処理の結果に示したガ格、ヲ格、ニ格、デ格の４つの格のそれぞれでの精度の平均での評価も行なった。この評価によれば、最も頻度の高い分類に決め打ちにする手法だと精度は２５％となる。併用型、借用型、非借用型ともに、この２５％の精度よりは高いことがわかる。

【０１４０】平均での評価では、精度の順は併用型、借用型、非借用型となっている。非借用型機械学習法は、問題に密接な教師信号を用いるために高い精度を得やすいとはいえ、本例のように事例数が少ない場合には他の機械学習法よりも精度が低くなる場合があることがわかる。

【０１４１】併用型機械学習法は、図９に示す評価においても、借用型機械学習法に１％劣っているだけで、図１０に示す平均での評価では圧倒的によく、両方の評価基準ともによい結果を得ている。

【０１４２】以上のことから、借用型機械学習法がランダムな選択より有効であり、かつ分類先の平均を評価基準とすると非借用型機械学習法より有効であることがわかる。また、併用型機械学習法が複数の評価基準で安定してよい結果を示したことがわかる。よって、借用型機械学習法と併用型機械学習法の有効性が示された。

【０１４３】次に、外の関係などの「その他」の分類も含めた格解析全般の処理を行なった。この処理では、評価用のデータ（１，５３０事例）をすべて用いた。この処理では併用型および非借用型の２つの機械学習法で行った。借用教師信号だけでは「その他」の分類を特定できないため、借用型機械学習法は用いなかった。図１１に、この処理の結果を示す。

【０１４４】また、この処理では、ガ格、ヲ格、ニ格、デ格、”その他”の５つの分類先のそれぞれでの精度の平均でも評価した。図１２に、この処理の結果を示す。処理結果から、サポートベクトルマシン法による処理の精度が最も良く、また、併用型機械学習法は全事例での処理の精度で１％ほど非借用より低いだけであって、平均精度では併用型機械学習法の方が圧倒的に高かった。

【０１４５】以上の具体例に示すように、教師信号借用型機械学習法がランダムな解析よりも精度が高くまた分類先ごとの精度を平均した精度では非借用型機械学習法

よりも精度が高いことがわかった。また、併用型機械学習法が全事例での精度だけでなく、分類先ごとの精度を平均した精度でも高く複数の評価基準において安定して高い精度を得ることを確認した。これらのことから、本発明の解析処理における有効性が確認された。

【０１４６】以上、本発明をその実施の態様により説明したが、本発明はその主旨の範囲において種々の変形が可能である。

【０１４７】

【発明の効果】以上説明したように、本発明によれば、従来の教師信号以外に大量の教師信号を借用することができるため、使用する教師信号が増加し、よって学習の精度向上が期待できる。

【０１４８】特に、本発明にかかる併用型機械学習法は、省略補完処理、文生成処理、機械翻訳処理、文字認識処理、音声認識処理など、語句を生成する処理を含むような極めて広範囲の問題に適用することができ、実用性の高い言語処理システムを実現することができる。

【０１４９】機械学習法には、さまざまな高度な手法が提案されている。本発明では、格解析などの言語処理を機械学習法の問題として扱うことができるように変換する。これにより、その時に応じた最もよい機械学習法を選択して言語処理の問題を解くことができる。

【０１５０】また、よりよい手法を用いることに加えて、より良い、かつ、より多くのデータ、素性を用いることは、処理精度の向上に必要である。本発明では、教師信号借用型機械学習法や併用型機械学習法を用いることにより、広範な情報を利用して解析に関係する広範な問題を取り扱うことができ、特に、教師信号借用型機械学習法により、人手で解析情報を付与していない用例を使用することができる。これにより、労力の負担を伴わずにより多くの情報を利用できることによる処理精度の向上を図ることができるという効果を奏する。

【０１５１】また、本発明では併用型機械学習法により、多くの情報を用いることに加えて、従来の教師信号を用いたより良い情報をも用いて言語処理を行う。これにより、いっそうの処理の精度の向上を図ることができるという効果を奏する。

【図面の簡単な説明】

【図１】本発明にかかるシステムの構成例を示す図である。

【図２】教師データの生成処理の処理フローチャートである。

【図３】教師信号借用型機械学習法による解析処理の処理フローチャートである。

【図４】機械学習法としてサポートベクトルマシン法を用いる場合のシステム構成例を示す図である。

【図５】サポートベクトルマシン法のマージン最大化の概念を示す図である。

【図６】機械学習法としてサポートベクトルマシン法を

25

用いた場合の解析処理の処理フローチャートである。

【図７】全事例における分類先の出現の分布を示す図である。

【図８】格助詞の再推定問題の処理の精度を示す図である。

【図９】主題化・連体化現象における表層格復元の処理の精度を示す図である。

【図１０】主題化・連体化現象における表層格復元の処理の精度の平均を示す図である。

【図１１】格解析全般での処理の精度を示す図である。

【図１２】格解析全般での処理の精度の平均を示す図である。

【符号の説明】

１　言語解析処理システム（ＣＰＵ／メモリ）

26

２　生コーパス

３　データ

４　解析情報

１１　問題表現相当部抽出部

１２　問題表現情報記憶部

１３　問題構造変換部

１４　意味解析情報記憶部

１５　教師データ記憶部

１６　解データベース

１７　解－素性対抽出部

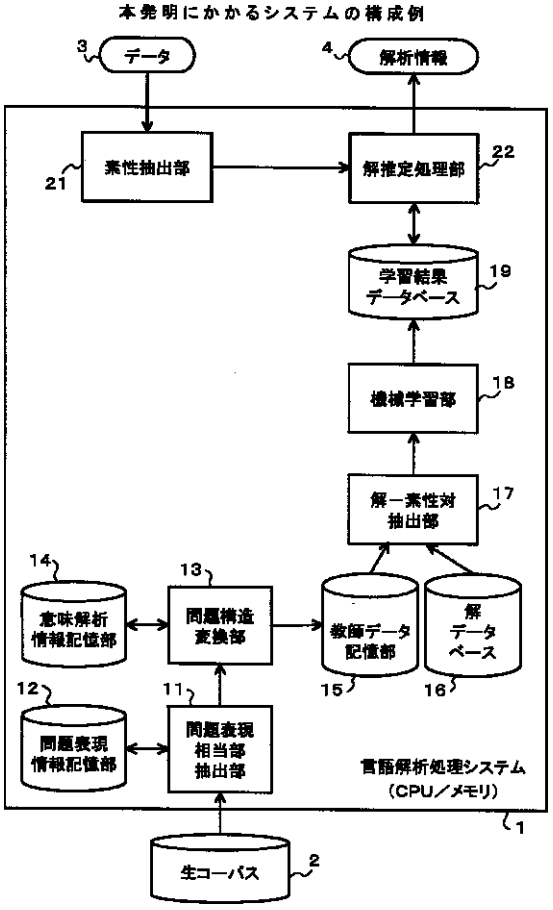
１８　機械学習部

１９　学習結果データベース

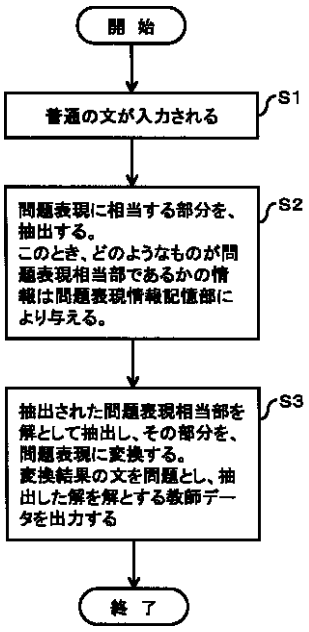
２１　素性抽出部

２２　解推定処理部

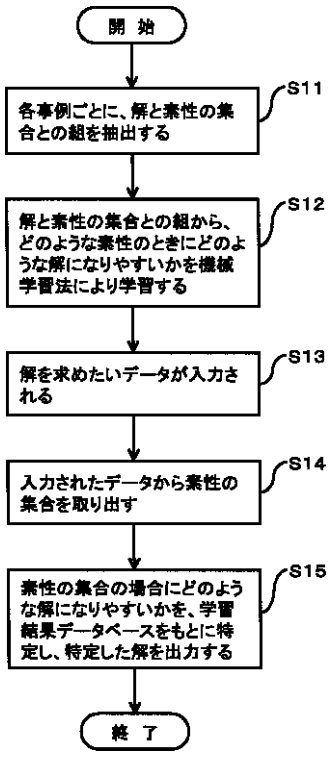
【図１】



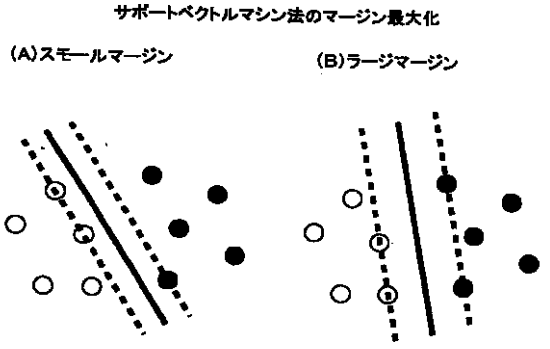
【図２】



【図３】



【図５】

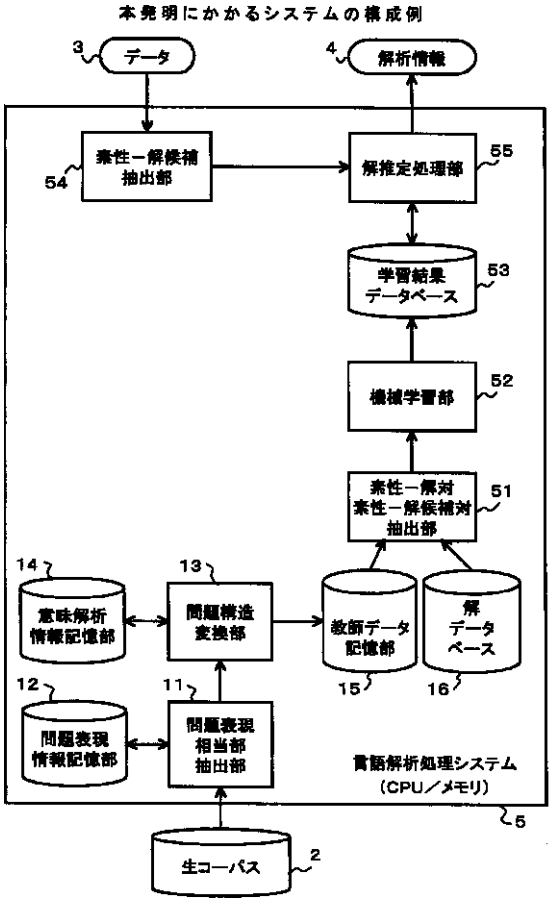


【図８】

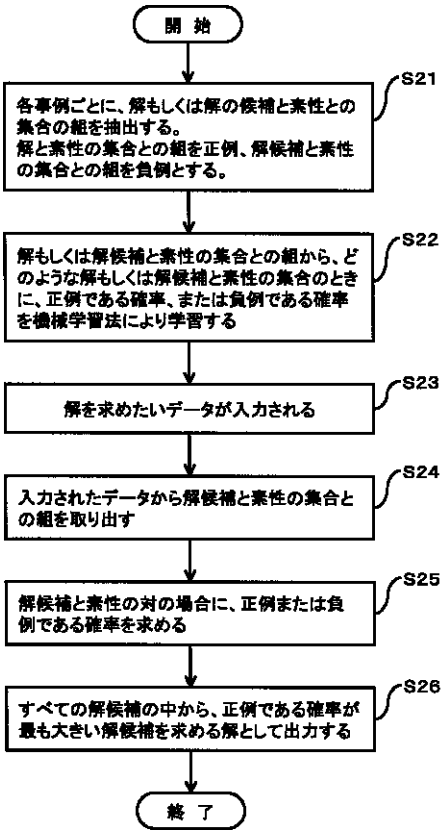
格助詞の再推定問題の処理の精度

TiMBL	SB	DL	ME	SVM
27. 40%	50. 22%	54. 70%	66. 93%	70. 25%

【図４】



【図６】



【図７】

分類先の出現の分布

	主題化	連体化
ガ格	526	499
ヲ格	29	46
ニ格	14	45
デ格	16	10
ト格	0	2
カラ格	0	1
その他	75	267
合計	660	870

【図９】

主題化・連体化現象における表層格復元の処理の精度

	TiMBL	SB	DL	ME	SVM
併用型	9. 85%	71. 13%	75. 34%	82. 24%	87. 04%
借用型	10. 61%	44. 11%	33. 42%	51. 18%	55. 39%
非借用型	86. 03%	82. 07%	84. 68%	86. 87%	88. 22%

【図１１】

格解析全般での処理の精度

	TiMBL	SB	DL	ME	SVM
併用型	8. 95%	65. 42%	51. 50%	71. 63%	81. 57%
非借用型	64. 05%	70. 00%	72. 35%	80. 46%	82. 55%

【図１０】

主題化・連体化現象における表層格復元の処理の精度の平均

	TiMBL	SB	DL	ME	SVM
併用型	21. 90%	45. 74%	37. 33%	51. 35%	62. 16%
借用型	28. 63%	48. 31%	31. 37%	54. 40%	59. 11%
非借用型	24. 93%	52. 79%	31. 95%	42. 90%	44. 96%

【図１２】

格解析全般での処理の精度の平均

	TIMBL	SB	DL	ME	SVM
併用型	24. 35%	43. 57%	29. 28%	46. 57%	56. 93%
非併用型	22. 90%	50. 23%	33. 67%	46. 29%	47. 03%