(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3780341号
(P3780341)
(45)発行日　平成18年5月31日(2006.5.31)　(24)登録日　平成18年3月17日(2006.3.17)

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

請求項の数 7　(全 47 頁)

(21)出願番号　特願2002-337747(P2002-337747)
(22)出願日　平成14年11月21日(2002.11.21)
(65)公開番号　特開2004-171354(P2004-171354A)
(43)公開日　平成16年6月17日(2004.6.17)
審査請求日　平成14年11月21日(2002.11.21)

特許法第３０条第１項適用　２００２年５月２３日　社団法人情報処理学会開催の「情報処理学会研究報告　Ｖｏｌ．２００２，Ｎｏ．４４　２００２－ＮＬ－１４９－６」において文書をもって発表

(73)特許権者　301022471
独立行政法人情報通信研究機構
東京都小金井市貫井北町４－２－１
(74)代理人　100119161
弁理士　重久　啓子
(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

審査官　和田　財太

最終頁に続く

(54)【発明の名称】言語解析処理システムおよび文変換処理システム

(57)【特許請求の範囲】
【請求項１】
機械学習処理を用いて言語解析処理を行うメイン用処理システムと、前記メイン用処理システムに対して機械学習処理で使用するデータを提供するスタック用処理システムとで構成され、所定の言語解析処理を行う言語解析処理システムであって、
前記スタック用処理システムは、
前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報を含まない文データを記憶する文データ記憶手段と、
前記問題が示される所定の文表現である問題表現と、前記問題表現に相当する部分とを組にして記憶する問題表現情報記憶手段と、
前記文データ記憶手段に記憶された文データから、前記問題表現に相当する部分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相当部抽出手段と、
前記文データの問題表現相当部を前記問題表現で変換した変換文を問題とし、前記問題表現相当部を解として、問題と解との組である教師なしデータを作成する問題構造変換手段と、
前記作成された教師なしデータを記憶する教師なしデータ記憶手段と、
前記教師なしデータ記憶手段に記憶された教師なしデータの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記教師なしデータごとに前記素性の集合と解との組を生成するスタック用解－素性対抽出手段と、

所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理し、学習結果として、前記どのような素性の集合との場合にどのような解になりやすいかということをスタック用学習結果データ記憶手段に保存するスタック用機械学習手段と、

前記メイン用処理システムから、前記スタック用解－素性対抽出手段が行う抽出処理と同様の抽出処理によって抽出された前記所定の情報である素性の集合を受け取った場合に、前記スタック用学習結果データ記憶手段に学習結果として記憶された前記どのような素性の集合の場合にどのような解になりやすいかということにもとづいて、前記素性の集合の場合になりやすい解を推定し、前記推定した解をスタック用出力解として出力するスタック用解推定処理手段とを備え、

前記メイン用処理システムは、

問題と解とで構成された文データであって、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段と、

前記解データ記憶手段に記憶された解データの問題から、前記スタック用解－素性対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解との組を生成するメイン用解－素性対抽出手段と、

前記メイン用解－素性対抽出手段で生成された前記素性の集合に対して前記スタック用解推定処理手段において推定され出力された前記スタック用出力解を、前記メイン用解－素性対抽出手段によって生成された素性の集合に素性として追加し、第１の素性の集合とする第１素性追加手段と、

所定の機械学習アルゴリズムにもとづいて、前記第１の素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理し、学習結果として、前記どのような素性の集合の場合にどのような解になりやすいかということをメイン用学習結果データ記憶手段に保存するメイン用機械学習手段と、

前記言語解析処理の対象として入力された入力文データから、前記スタック用解－素性対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性として抽出する素性抽出手段と、

前記素性抽出手段で生成された前記素性の集合に対して前記スタック用解推定処理手段において推定され出力されたスタック用出力解を、前記素性抽出手段によって生成された素性の集合に素性として追加し、第２の素性の集合とする第２素性追加手段と、

前記メイン用学習結果データ記憶手段に学習結果として記憶された前記どのような素性の集合の場合にどのような解になりやすいかということにもとづいて、前記第２の素性の集合の場合になりやすい解を推定する解推定処理手段とを備え、

前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、

前記決定リスト法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記最大エントロピー法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記サポートベクトルマシン法では、前記スタック用機械学習手段および前記メイン用

機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推定される処理が行われる

ことを特徴とする言語解析処理システム。

【請求項２】

前記スタック用処理システムは、問題と解とで構成され、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段を備えるとともに、

前記スタック用解－素性対抽出手段は、前記解データ記憶手段に記憶された解データの問題から、前記抽出処理によって前記所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解との組を生成し、

前記スタック用機械学習手段は、前記文データおよび前記解データから生成された素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理する

ことを特徴とする請求項１記載の言語解析処理システム。

【請求項３】

機械学習処理を用いて言語解析処理を行うメイン用処理システムと、前記メイン用処理システムに対して機械学習処理で使用するデータを提供するスタック用処理システムとで構成され、所定の言語解析処理を行う言語解析処理システムであって、

前記スタック用処理システムは、

前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報を含まない文データを記憶する文データ記憶手段と、

前記問題が示される所定の文表現である問題表現と、前記問題表現に相当する部分とを組にして記憶する問題表現情報記憶手段と、

前記文データ記憶手段に記憶された文データから、前記問題表現に相当する部分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相当部抽出手段と、

前記文データの問題表現相当部を前記問題表現で変換した変換文を問題とし、前記問題表現相当部を解または解候補として、問題と解または解候補との組である教師なしデータを作成する問題構造変換手段と、

前記作成された教師なしデータを記憶する教師なしデータ記憶手段と、

前記教師なしデータ記憶手段に記憶された教師なしデータの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記教師なしデータごとに前記素性の集合と解または解候補との組を生成するスタック用素性－解対・素性－解候補対抽出手段と、

所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解または解候補との組について、どのような素性の集合と解または解候補との組の場合に所定の二分類先である正例もしくは負例である確率を機械学習処理し、学習結果として、前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率をスタック用学習結果データ記憶手段に保存するスタック用機械学習手段と、

前記メイン用処理システムから、前記スタック用素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって抽出された前記所定の情報である素性とする素性の集合と解または解候補との組を受け取った場合に、前記学習結果データ記憶手段に学習結果として記憶された前記素性の集合と解または解候補の組の場合に正例もしくは負例である確率にもとづいて、前記素性の集合と解候補との組の場合に正例もしくは負例である確率を求め、全ての解候補の中から正例である確率が最大の解候補をスタック用出力解として出力するスタック用解推定処理手段とを備え、

前記メイン用処理システムは、
問題と解とで構成された文データであって、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段と、
前記解データ記憶手段に記憶された解データの問題から、前記スタック用素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性を抽出し、前記素性の集合と前記解または解候補との組を生成するメイン用素性－解対・素性－解候補対抽出手段と、
前記メイン用素性－解対・素性－解候補対抽出手段で生成された前記素性の集合と解または解候補との組に対して前記スタック用解推定処理手段において推定され出力されたスタック用出力解を、前記メイン用解－素性対抽出手段によって生成された素性の集合に素性として追加し、第１の素性の集合とする第１素性追加手段と、
所定の機械学習アルゴリズムにもとづいて、前記解と第１の素性の集合と解または解候補との組について、前記素性の集合と解または解候補の場合に正例もしくは負例である確率を機械学習処理し、学習結果として、前記素性の集合と解または解候補の場合に正例もしくは負例である確率をメイン用学習結果データ記憶手段に保存するメイン用機械学習手段と、
前記言語解析処理の対象として入力された入力文データから、前記スタック用素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性として抽出する素性抽出手段と、
前記素性抽出手段で生成された前記素性の集合と解または解候補の組に対して前記スタック用解推定処理手段において推定され出力されたスタック用出力解を、前記素性抽出手段によって生成された素性の集合に素性として追加し、第２の素性の集合とする第２素性追加手段と、
前記メイン用学習結果データ記憶手段に学習結果として記憶された前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率にもとづいて、前記第２の素性の集合と解候補との組の場合に正例もしくは負例である確率を求め、全ての解候補の中から正例である確率が最大の解候補を解として推定する解推定処理手段とを備え、
前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、
前記決定リスト法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、
前記最大エントロピー法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、
前記サポートベクトルマシン法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推

定される処理が行われる
ことを特徴とする言語解析処理システム。
【請求項４】
前記スタック用処理システムは、問題と解とで構成され、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段を備えるとともに、
前記スタック用解－素性対抽出手段は、前記解データ記憶手段に記憶された解データの問題から、前記抽出処理によって前記所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解との組を生成し、
前記スタック用機械学習手段は、前記文データおよび前記解データから生成された素性の集合と解または解候補との組について、前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率を機械学習処理する
ことを特徴とする請求項３記載の言語解析処理システム。
【請求項５】
前記スタック用処理システムおよび前記メイン用処理システムでは、前記言語解析処理の対象となる文データが受け身文または使役文である場合に、前記文データから能動文への文変換処理における変換後の格助詞を解析する
ことを特徴とする請求項１ないし請求項４のいずれか一項に記載の言語解析処理システム。
【請求項６】
機械学習処理を用いて、受け身文または使役文である文データを能動文の文データへ変換する場合の変換後の格助詞を推定する文変換処理システムであって、
問題と解とで構成されたデータであって、文データを問題とし、前記変換処理での問題に対する解情報を解とする解データを記憶する解データ記憶手段と、
前記解データ記憶手段に記憶された解データの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解との組を生成する解－素性対抽出手段と、
所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理し、学習結果として、前記どのような素性の集合の場合にどのような解になりやすいかということを学習結果データ記憶手段に保存する機械学習手段と、
前記変換処理の対象として入力された入力文データから、前記解－素性対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性として抽出する素性抽出手段と、
前記学習結果データ記憶手段に学習結果として記憶された前記どのような素性の集合の場合にどのような解になりやすいかということにもとづいて、前記素性の集合の場合になりやすい解を推定する解推定処理手段とを備え、
前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、
前記決定リスト法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、
前記最大エントロピー法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集

合のときになりやすい解として推定される処理が、または、
前記サポートベクトルマシン法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推定される処理が行われる
ことを特徴とする文変換処理システム。
【請求項７】
機械学習処理を用いて、受け身文または使役文である文データを能動文の文データへ変換する場合の変換後の格助詞を推定する文変換処理システムであって、
問題と解とで構成されたデータであって、文データを問題とし、前記変換処理での問題に対する解情報を解とする解データを記憶する解データ記憶手段と、
前記解データ記憶手段に記憶された前記解データの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解または解候補との組を生成する素性－解対・素性－解候補対抽出手段と、
所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解または解候補との組について、どのような素性の集合と解または解候補との組の場合に正例もしくは負例である確率を機械学習処理し、学習結果として、前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率を学習結果データ記憶手段に保存する機械学習手段と、
前記変換処理の対象として入力された入力文データから、前記素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性を抽出し、前記素性の集合と解候補との組を生成する素性－解候補対抽出手段と、
前記学習結果データ記憶手段に学習結果として記憶された前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率にもとづいて、前記素性の集合と解候補との組の場合に正例もしくは負例である確率を求め、全ての解候補の中から正例である確率が最大の解候補を解として推定する解推定処理手段とを備え、
前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、
前記決定リスト法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、
前記最大エントロピー法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、
前記サポートベクトルマシン法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推定される処理が行われる

ことを特徴とする文変換処理システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】
本発明は、コンピュータで実現する自然言語処理技術に関する。さらに詳しくは、機械学習法により電子化された文を用いた言語解析処理方法および前記処理方法を実現する処理システムに関する。
【０００２】
特に、本発明は、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な問題を扱う言語処理に適用することができる。
【０００３】
【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。
【０００４】
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である。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。このように、「りんごは食べた。」の「りんごは」の「は」の部分を「ヲ格」と解析する処理である。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。この場合に、「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。例えば、「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】
このような言語解析処理をコンピュータで実現する場合に、処理を行う者の労力の負担を軽減しつつ高い処理精度を得るために、機械学習法を用いて言語解析処理を行う手法を提示した（非特許文献１参照）。
【０００７】
非特許文献１において提示した機械学習法を用いて言語解析処理を行う手法（非借用型機械学習法）は、以下のような利点を備える。
(i) より大きな教師データを持つコーパスを用意することで、さらに高い精度で処理を行えることができると推測できる。
(ii)よりよい機械学習手法が開発されたとき、その機械学習手法を用いることでさらに高い精度を獲得できると予測できる。
【０００８】
さらに、非特許文献１では、借用型機械学習法を用いた言語解析処理方法を提示した。借用型機械学習法とは、機械学習法の解析対象となる情報が付加されていないデータ（以下「教師なしデータ」という。）から生成した教師信号を用いた機械学習方法である。借用型機械学習法によれば、例えば格フレーム辞書など、人手などで解析対象となる情報（解情報）を予め付与しておいたデータを用いることなく、大量に存在する一般的な電子化された文を機械学習の教師なしデータとして利用することができ、大量の教師信号による機械学習の学習精度が向上するため、高い精度の言語解析処理を実現することができる。
【０００９】
さらに、非特許文献１では、併用型機械学習法を用いた言語解析処理方法を提示した。併用型機械学習法とは、通常の機械学習法で用いる教師信号すなわち機械学習法の解析対象

となる情報が付加されたデータ（以下「教師ありデータ」という。）と、教師なしデータから生成した教師信号とを用いて機械学習を行う方法である。併用型機械学習法によれば、取得が容易な教師なしデータから生成された大量の教師信号と、通常の学習精度を確保できる教師ありデータの教師信号との両方の利点を活かした言語解析処理を実現することができる。

【００１０】

また、自然言語処理の分野における重要な問題として、受け身文や使役文から能動文への変換処理がある。この文変換処理は、文生成処理、言い換え処理、文の平易化／言語運用支援、自然言語文を利用した知識獲得・情報抽出処理、質問応答システムなど、多くの研究分野で役に立つ。例えば質問応答システムにおいて、質問文が能動文で書かれ回答を含む文が受動文で書かれているような文書がある場合に、質問文と回答を含む文では文構造が異なっているために質問の回答を取り出すのが困難な場合がある。このような問題も、受け身文や使役文から能動文への変換処理を行うことにより解決することができる。

【００１１】

日本語の受け身文や使役文を能動文に文変換処理する際には、文変換後に用いる変換後格助詞を推定することが求められる。例えば、「犬に私が噛まれた。」という受け身文から「犬が私を噛んだ。」という能動文に変換する場合に、「犬に」の格助詞「に」が「が」に、「私が」の「が」が「を」に変換されると推定する処理である。また、「彼が彼女に髪を切らせた。」という使役文を「彼女が髪を切った。」という能動文に変換する場合に、「彼女に」の格助詞「に」が「が」に変換され、「髪を」の「を」は変換しないと推定する処理である。しかし、受け身文や使役文から能動文への変換処理における格助詞の変換は、変換される格助詞が動詞やその動詞の使われ方に依存して変わるので、簡単に自動処理できる問題ではない。

【００１２】

格助詞の変換処理については、例えば、以下の非特許文献２～４に示すような従来手法がいくつかある。非特許文献２～４で開示されている技術では、格助詞の変換処理の問題を、どのように格助詞を変換すればよいかを記載した格フレーム辞書を用いて対処している。

【００１３】

【非特許文献１】

村田真樹、
機械学習手法を用いた日本語格解析－教師信号借用型と非借用型さらには併用型－、
電子情報通信学会、電子情報通信学会技術研究報告NLC-2001-24
２００１年７月１７日

【非特許文献２】

情報処理振興事業協会技術センター、
計算機用日本語基本動詞辞書ＩＰＡＬ(Basic Verbs) 説明書、
１９８７

【非特許文献３】

Sadao Kurohashi and Makoto Nagao,
A Method of Case Structure Analysis for Japanese Sentences based on Examples in Case Frame Dictionary,
IEICE Transactions of Information and Systems, Vol.E77-D, No.2, １９９４

【非特許文献４】

近藤　恵子、佐藤　理史、奥村　学、
格変換による単文の言い換え、
情報処理学会論文誌、Vol.42, No.3,
２００１

【００１４】

【発明が解決しようとする課題】

前記の非特許文献１は、機械学習法を言語解析処理に適用することで処理精度を向上させるという効果を持つ。また、借用型機械学習法や併用型機械学習法は、人手による労力負担を増やすことなく機械学習の教師信号を増大させることができる点で非常に有効である。

【００１５】

機械学習処理では、与えられた教師データにおいて正解率を最大とするように学習を行う。また、教師なしデータは、解析対象となる情報を持たないという点で教師ありデータと異なる性質のものである。

【００１６】

したがって、非特許文献１に示す併用型機械学習法のように単純に教師なしデータを教師ありデータに追加した教師信号を用いた機械学習処理は、教師ありデータと教師なしデータとを合計したデータでの正解率を最大にするように学習する。そのため、教師なしデータと教師ありデータとの関係によっては教師ありデータだけでの正解率を最大にするように学習した機械学習の場合に比べて学習精度が低下してしまうという問題が生ずる。

【００１７】

このような従来技術の問題に鑑みると、教師ありデータと教師なしデータの利点を活かして、より確実に精度の高い学習処理が行えるような手法の実現が求められる。

【００１８】

また、受け身文・使役文から能動文への文変換処理について、前記の非特許文献２～４に示すような従来の技術では、どのように格助詞を変換すればよいかをすべての動詞とその動詞の使い方について記載した格フレーム辞書が必要であった。

【００１９】

しかし、すべての動詞とその動詞の使い方を記載した辞書を用意することは事実上困難であるため、この格フレーム辞書を用いた変換処理方法は不十分であり、格フレーム辞書に記載されていない動詞や動詞の使い方がされた文を変換することができなかったり、誤変換する確率が高かったりするという問題が生じていた。

【００２０】

したがって、特に受け身文・使役文から能動文への文変換処理について、人手による労力負担を増大させずに高い精度の処理が行えるような手法が求められる。

【００２１】

本発明の目的は、教師ありデータと教師なしデータの両方を用いて機械学習を行う併用型教師学習法を用いて言語解析処理を行う場合に、双方のデータの利点を活かして、より高い精度で言語解析処理を行える処理システムを提供することである。

【００２２】

さらに、本発明の目的は、特に受け身文や使役文から能動文への文変換処理について、機械学習法を用いて高い精度で変換後格助詞を推定できる文変換処理システムを提供することである。

【００２３】

【課題を解決するための手段】

上記の目的を達成するため、本発明は以下のような構成をとる。

【００２４】

本発明は、機械学習処理を用いて言語解析処理を行うメイン用処理システムと、前記メイン用処理システムに対して機械学習処理で使用するデータを提供するスタック用処理システムとで構成され、所定の言語解析処理を行う言語解析処理システムであって、

前記スタック用処理システムは、１）前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報を含まない文データを記憶する文データ記憶手段と、

前記問題が示される所定の文表現である問題表現と、前記問題表現に相当する部分とを組にして記憶する問題表現情報記憶手段と、２）前記文データ記憶手段に記憶された文データから、前記問題表現に相当する部分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相当部抽出手段と、３）前記文データの問題表現相当部を前記問題表現で変換し

た変換文を問題とし、前記問題表現相当部を解として、問題と解との組である教師なしデータを作成する問題構造変換手段と、４）前記作成された教師なしデータを記憶する教師なしデータ記憶手段と、５）前記教師なしデータ記憶手段に記憶された教師なしデータの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記教師なしデータごとに前記素性の集合と解との組を生成するスタック用解－素性対抽出手段と、６）所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理し、学習結果として、前記どのような素性の集合との場合にどのような解になりやすいかということをスタック用学習結果データ記憶手段に保存するスタック用機械学習手段と、７）前記メイン用処理システムから、前記スタック用解－素性対抽出手段が行う抽出処理と同様の抽出処理によって抽出された前記所定の情報である素性の集合を受け取った場合に、前記スタック用学習結果データ記憶手段に学習結果として記憶された前記どのような素性の集合の場合にどのような解になりやすいかということにもとづいて、前記素性の集合の場合になりやすい解を推定し、前記推定した解をスタック用出力解として出力するスタック用解推定処理手段とを備え、

前記メイン用処理システムは、８）問題と解とで構成された文データであって、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段と、９）前記解データ記憶手段に記憶された解データの問題から、前記スタック用解－素性対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解との組を生成するメイン用解－素性対抽出手段と、１０）前記メイン用解－素性対抽出手段で生成された前記素性の集合に対して前記スタック用解推定処理手段において推定され出力された前記スタック用出力解を、前記メイン用解－素性対抽出手段によって生成された素性の集合に素性として追加し、第１の素性の集合とする第１素性追加手段と、１１）所定の機械学習アルゴリズムにもとづいて、前記第１の素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理し、学習結果として、前記どのような素性の集合の場合にどのような解になりやすいかということをメイン用学習結果データ記憶手段に保存するメイン用機械学習手段と、１２）前記言語解析処理の対象として入力された入力文データから、前記スタック用解－素性対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性として抽出する素性抽出手段と、１３）前記素性抽出手段で生成された前記素性の集合に対して前記スタック用解推定処理手段において推定され出力されたスタック用出力解を、前記素性抽出手段によって生成された素性の集合に素性として追加し、第２の素性の集合とする第２素性追加手段と、１４）前記メイン用学習結果データ記憶手段に学習結果として記憶された前記どのような素性の集合の場合にどのような解になりやすいかということにもとづいて、前記第２の素性の集合の場合になりやすい解を推定する解推定処理手段とを備え、

前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、

前記決定リスト法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記最大エントロピー法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集合のと

きになりやすい解として推定される処理が、または、
前記サポートベクトルマシン法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推定される処理が行われることを特徴とする。
【００２５】
また、前記スタック用処理システムは、１５）問題と解とで構成され、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段を備えるとともに、
前記スタック用解－素性対抽出手段は、前記解データ記憶手段に記憶された解データの問題から、前記抽出処理によって前記所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解との組を生成し、前記スタック用機械学習手段は、前記文データおよび前記解データから生成された素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理することを特徴とする。
【００２６】
さらに、本発明は、機械学習処理を用いて言語解析処理を行うメイン用処理システムと、前記メイン用処理システムに対して機械学習処理で使用するデータを提供するスタック用処理システムとで構成され、所定の言語解析処理を行う言語解析処理システムであって、
前記スタック用処理システムは、１）前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報を含まない文データを記憶する文データ記憶手段と、２）前記問題が示される所定の文表現である問題表現と、前記問題表現に相当する部分とを組にして記憶する問題表現情報記憶手段と、３）前記文データ記憶手段に記憶された文データから、前記問題表現に相当する部分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相当部抽出手段と、４）前記文データの問題表現相当部を前記問題表現で変換した変換文を問題とし、前記問題表現相当部を解または解候補として、問題と解または解候補との組である教師なしデータを作成する問題構造変換手段と、５）前記作成された教師なしデータを記憶する教師なしデータ記憶手段と、６）前記教師なしデータ記憶手段に記憶された教師なしデータの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記教師なしデータごとに前記素性の集合と解または解候補との組を生成するスタック用素性－解対・素性－解候補対抽出手段と、７）所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解または解候補との組について、どのような素性の集合と解または解候補との組の場合に所定の二分類先である正例もしくは負例である確率を機械学習処理し、学習結果として、前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率をスタック用学習結果データ記憶手段に保存するスタック用機械学習手段と、８）前記メイン用処理システムから、前記スタック用素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって抽出された前記所定の情報である素性とする素性の集合と解または解候補との組を受け取った場合に、前記学習結果データ記憶手段に学習結果として記憶された前記素性の集合と解または解候補の組の場合に正例もしくは負例である確率にもとづいて、前記素性の集合と解候補との組の場合に正例もしくは負例である確率を求め、全ての解候補の中から正例である確率が最大の解候補をスタック用出力解として出力するスタック用解推定処理手段とを備え、
前記メイン用処理システムは、９）問題と解とで構成された文データであって、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段と、１０）前記解データ記憶手段に記憶された解

データの問題から、前記スタック用素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性を抽出し、前記素性の集合と前記解または解候補との組を生成するメイン用素性－解対・素性－解候補対抽出手段と、１１）前記メイン用素性－解対・素性－解候補対抽出手段で生成された前記素性の集合と解または解候補との組に対して前記スタック用解推定処理手段において推定され出力されたスタック用出力解を、前記メイン用解－素性対抽出手段によって生成された素性の集合に素性として追加し、第１の素性の集合とする第１素性追加手段と、１２）所定の機械学習アルゴリズムにもとづいて、前記解と第１の素性の集合と解または解候補との組について、前記素性の集合と解または解候補の場合に正例もしくは負例である確率を機械学習処理し、学習結果として、前記素性の集合と解または解候補の場合に正例もしくは負例である確率をメイン用学習結果データ記憶手段に保存するメイン用機械学習手段と、１３）前記言語解析処理の対象として入力された入力文データから、前記スタック用素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性として抽出する素性抽出手段と、１４）前記素性抽出手段で生成された前記素性の集合と解または解候補の組に対して前記スタック用解推定処理手段において推定され出力されたスタック用出力解を、前記素性抽出手段によって生成された素性の集合に素性として追加し、第２の素性の集合とする第２素性追加手段と、１５）前記メイン用学習結果データ記憶手段に学習結果として記憶された前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率にもとづいて、前記第２の素性の集合と解候補との組の場合に正例もしくは負例である確率を求め、全ての解候補の中から正例である確率が最大の解候補を解として推定する解推定処理手段とを備え、

前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、

前記決定リスト法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記最大エントロピー法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記サポートベクトルマシン法では、前記スタック用機械学習手段および前記メイン用機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記スタック用解推定処理手段および前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推定される処理が行われることを特徴とする。

【００２７】

また、前記スタック用処理システムは、１６）問題と解とで構成され、前記言語解析処理での解析対象であって機械学習処理で扱われる問題に対する解情報が付与された解データを記憶する解データ記憶手段を備えるとともに、

前記スタック用解－素性対抽出手段は、前記解データ記憶手段に記憶された解データの問題から、前記抽出処理によって前記所定の情報である素性を抽出し、前記解データごと

に前記素性の集合と解との組を生成し、前記スタック用機械学習手段は、前記文データおよび前記解データから生成された素性の集合と解または解候補との組について、前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率を機械学習処理することを特徴とする。

【００２８】

このように、本発明では、教師なしデータを用いた機械学習法による解析結果を教師ありデータの素性として組み込むことにより、機械学習処理において教師ありデータについての正解率を最大とするように学習が行われるため、異なる性質の教師なしデータと教師ありデータとの双方の利点を活かした機械学習処理を行うことができ、高い精度の解析処理を実現することができる。

【００２９】

さらに、本発明は、機械学習処理を用いて、受け身文または使役文である文データを能動文の文データへ変換する場合の変換後の格助詞を推定する文変換処理システムであって、１）問題と解とで構成されたデータであって、文データを問題とし、前記変換処理での問題に対する解情報を解とする解データを記憶する解データ記憶手段と、２）前記解データ記憶手段に記憶された解データの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解との組を生成する解－素性対抽出手段と、３）所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解との組について、どのような素性の集合の場合にどのような解になりやすいかということを機械学習処理し、学習結果として、前記どのような素性の集合の場合にどのような解になりやすいかということを学習結果データ記憶手段に保存する機械学習手段と、４）前記変換処理の対象として入力された入力文データから、前記解－素性対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性として抽出する素性抽出手段と、５）前記学習結果データ記憶手段に学習結果として記憶された前記どのような素性の集合の場合にどのような解になりやすいかということにもとづいて、前記素性の集合の場合になりやすい解を推定する解推定処理手段とを備え、

前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、

前記決定リスト法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記最大エントロピー法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記サポートベクトルマシン法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推定される処理が行われることを特徴とする。

【００３０】

さらに、本発明は、機械学習処理を用いて、受け身文または使役文である文データを能

動文の文データへ変換する場合の変換後の格助詞を推定する文変換処理システムであって、１）問題と解とで構成されたデータであって、文データを問題とし、前記変換処理での問題に対する解情報を解とする解データを記憶する解データ記憶手段と、２）前記解データ記憶手段に記憶された前記解データの問題から、所定の解析処理によって、少なくとも文字列または単語または品詞を含む所定の情報である素性を抽出し、前記解データごとに前記素性の集合と解または解候補との組を生成する素性－解対・素性－解候補対抽出手段と、３）所定の機械学習アルゴリズムにもとづいて、前記素性の集合と解または解候補との組について、どのような素性の集合と解または解候補との組の場合に正例もしくは負例である確率を機械学習処理し、学習結果として、前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率を学習結果データ記憶手段に保存する機械学習手段と、４）前記変換処理の対象として入力された入力文データから、前記素性－解対・素性－解候補対抽出手段が行う抽出処理と同様の抽出処理によって前記所定の情報である素性を抽出し、前記素性の集合と解候補との組を生成する素性－解候補対抽出手段と、５）前記学習結果データ記憶手段に学習結果として記憶された前記素性の集合と解または解候補との組の場合に正例もしくは負例である確率にもとづいて、前記素性の集合と解候補との組の場合に正例もしくは負例である確率を求め、全ての解候補の中から正例である確率が最大の解候補を解として推定する解推定処理手段とを備え、

前記所定の機械学習アルゴリズムとして決定リスト法または最大エントロピー法またはサポートベクトルマシン法のいずれかのアルゴリズムを使用し、

前記決定リスト法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を規則とし、前記規則を所定の優先順位により格納したリストが前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果であるリストに格納された規則を優先順位の高い順に前記入力データの素性の集合と比較し、素性が一致した規則の解が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記最大エントロピー法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組から、前記素性の集合が所定の条件式を満足しかつエントロピーを示す式を最大にするときの確率分布が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である確率分布をもとに、前記入力データの素性の集合の場合の各分類の確率が求められ、前記確率が最大の確率値を持つ分類が、前記入力データの素性の集合のときになりやすい解として推定される処理が、または、

前記サポートベクトルマシン法では、前記機械学習手段によって、前記教師なしデータの素性の集合と解との組を用いて、所定のサポートベクトルマシン法による超平面を求め、前記超平面および前記超平面により分割された空間の分類が前記学習結果として記憶され、前記解推定処理手段によって、前記学習結果である超平面をもとに、前記入力文データの素性の集合が前記超平面で分割された空間のいずれかに属するかが求められ、前記素性の集合が属する空間の分類が、前記入力文データの素性の集合の場合になりやすい解として推定される処理が行われることを特徴とする。

【００３１】

受け身文や使役文から能動文への文変換処理における格助詞変換処理は、変換後の文で用いられる格助詞を決定することである。そして、変換後の格助詞の種類数は有限であるから、変換後の格助詞の推定問題は分類問題に帰着でき、機械学習手法を用いた処理として扱うことが可能である。

【００３２】

本発明では、解析対象についての情報（変換後格助詞など）を付与されていない文から生成されたデータ（教師なしデータ）を教師信号として機械学習を行う。これにより、大量に存在する通常の電子データ（文）を教師データとして利用することができ、解析対象についての情報を人手などにより付与するという労力負担を増加させることなく、高い精度の文変換処理を実現することができる。

【００３３】

【発明の実施の形態】
以下に本発明の実施の形態のいくつかを説明する。
【００３４】
第１の実施の形態として、受け身文・使役文から能動文への文変換処理に教師ありデータを用いた機械学習法（非借用型機械学習法）を適用する処理について説明する。また、第２の実施の形態として、受け身文・使役文から能動文への文変換処理に教師なしデータを用いた機械学習法（借用型機械学習法）を適用する処理について説明する。また、第３の実施の形態として、受け身文・使役文から能動文への文変換処理に教師ありデータと教師なしデータを併用して用いた機械学習法（併用型機械学習法）を適用する処理について説明する。
【００３５】
さらに、第４の実施の形態として、言語解析処理に、教師なしデータを用いた機械学習の結果を、教師ありデータの素性として用いた機械学習法（教師なしデータスタック型機械学習法）を適用する処理について説明する。
【００３６】
なお、本発明の実施の形態において、受け身文・使役文から能動文への変換処理での格助詞の変換処理とは、元の受け身文・使役文の格助詞を変換後の能動文の格助詞へ変換する処理、および元の受け身文・使役文の不要部分を消去する処理をいう。不要部分とは、使役文「彼が彼女に髪を切らせた。」から能動文「彼女が髪を切った。」への文変換において、元の使役文「彼が」の部分である。また、元の文（受け身文・使役文）の格助詞を変換前格助詞とし、能動文への文変換時に付与される新たな格助詞を変換後格助詞とする。
【００３７】
本形態では、これらの格助詞変換処理のみを対象にし、能動文への変換に伴う助動詞表現の変換処理などは処理対象として説明しない。助動詞表現部分程度の変換処理は、既存の処理、例えば文法に従った規則を用いる処理を用いて容易に実現することが可能である。
【００３８】
〔第１の実施の形態〕
第１の実施の形態として、受け身文・使役文から能動文への文変換処理を行う場合に、教師ありデータを用いた機械学習により、変更されるべき格助詞を自動変換処理する文変換処理システムの処理を説明する。
【００３９】
図１に、本形態における文変換処理システムの構成例を示す。文変換処理システム１００は、ＣＰＵおよびメモリからなり、解－素性対抽出部１０１、機械学習部１０２、学習結果データベース１０３、素性抽出部１１０、解推定処理部１１１および解データベース２を備える。
【００４０】
解－素性対抽出部１０１は、解データベース２から教師ありデータである事例を取り出し、事例ごとに事例の解と素性の集合との組（対）を抽出する手段である。
【００４１】
機械学習部１０２は、抽出された解と素性の集合との組から、どのうような素性のときにどのような解と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１０３に記憶する手段である。
【００４２】
素性抽出部１１０は、入力された文（受け身文または使役文）３から、素性の集合を抽出する手段である。なお、文は、文または少なくとも体言と用言を持つ文の一部とする。
【００４３】
解推定処理部１１１は、学習結果データベース１０３を参照して、入力文３の素性の場合にどのような解になりやすいか、すなわち能動文へ変換する場合に変換後格助詞になりやすい格助詞を推定し、推定した格助詞を解４として出力する手段である。
【００４４】

解データベース２は、機械学習で解析対象となる情報が付与された「問題－解」という構造を持つ教師ありデータを記憶する。本形態では、受け身文・使役文から能動文への変換処理における変換後格助詞が解析対象であり、能動文への変換処理で変更されるべき格助詞（変換後格助詞）の情報がタグ付けされた事例（単文）が記憶されたデータベースを利用することができる。

【００４５】

図２に、文変換処理システム１００の処理フローを示す。

【００４６】

ステップＳ１：　解－素性対抽出部１０１により、解データベース２から事例を取り出し、各事例ごとに解と素性の集合との組を抽出する。例えば、解データベース２として、受け身文や使役文のそれぞれの格助詞に対してそれが能動文になったときに用いられる変換後格助詞がタグとして付与されているタグ付きコーパスを用いる。

【００４７】

図３に、タグ付きコーパスに記憶されている事例（単文）を示す。図３に示す単文に下線を付けた５つの格助詞は変換前格助詞であり、下線部の下に矢印で示す格助詞は変換後格助詞を示す情報である。図３（Ａ）の事例は、この受け身文が能動文に変換される場合に、変換前格助詞が、それぞれ、「に」から「が」へ、「が」から「を」へ変換されることを意味する。また、図３（Ｂ）の事例は、この使役文が能動文に変換される場合に、変換前格助詞が、それぞれ、「に」から「が」へ、「を」から「を」へ変換され、「彼が」の部分は消去されることを意味している。「ｏｔｈｅｒ」は、その部分は能動文になるとき消去されることを意味するタグとする。

【００４８】

ここで、素性とは、機械学習法による解析処理で用いる細かい情報の１単位を意味する。抽出する素性としては、例えば以下のようなものがある。

【００４９】

１．体言ｎについている格助詞（変換前格助詞）
２．用言ｖの品詞
３．用言ｖの単語の基本形
４．用言ｖにつく助動詞列（例：「れる」、「させる」など）
５．体言ｎの単語
６．体言ｎの単語の分類語彙表の分類番号
７．用言ｖにかかる体言ｎ以外の体言がとる格

例えば、事例の問題が「犬に噛まれた。」である場合に、
・推定すべき格にある体言ｎの単語＝犬、
・推定すべき格が修飾する用言ｖ（単語の基本形）＝噛む、
・体言ｎと用言ｖとの間の格助詞（変換前格助詞）＝に、
などの素性が抽出される。

【００５０】

また、解は、各事例にタグ情報として付与された変換後格助詞であり、上記のの事例では、
・解（変換後格助詞）＝が
である。そして、解－素性対抽出部１０１は、抽出した素性の集合を機械学習部１０２で実行する機械学習処理での文脈とし、解を分類先とする。

【００５１】

ステップＳ２：　機械学習部１０２により、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法により学習し、この学習結果を学習結果データベース１０３に記憶する。

【００５２】

例えば、事例「犬に噛まれた。⇒が」から抽出された、
・推定すべき格にある体言ｎの単語＝犬、

・推定すべき格が修飾する用言ｖ（単語の基本形）＝噛む、
・体言ｎと用言ｖとの間の格助詞（変換前格助詞）＝に、
のような素性の集合の場合には、
・解（変換後格助詞）＝が
となりやすいことを学習する。

【００５３】

また、事例「ヘビに噛まれた。⇒が」から抽出された、
・推定すべき格にある体言ｎの単語＝ヘビ、
・推定すべき格が修飾する用言ｖ（単語の基本形）＝噛む、
・体言ｎと用言ｖとの間の格助詞（変換前格助詞）＝に、
のような素性の集合の場合にも、
・解（変換後格助詞）＝が
となりやすいことを学習する。

【００５４】

機械学習法は、例えば、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。

【００５５】

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ておき、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。

【００５６】

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（$1 \leqq j \leqq k$）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解（求める分類）とする方法である。

［参考文献１：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）］

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。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る。このため、通常、サポートベクトルマシン法にペアワイズ手法を組み合わせて使用することで、分類数が３個以上のデータを扱うことができる。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。

［参考文献２：Nello Cristianini and John Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods,(Cambridge University Press,2000) ］

［参考文献３：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を説明するため、図４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４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４（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４（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。

【００５７】

サポートベクトルマシン法の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４（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。

【００５８】

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など）がなされたものが用いられる。

【００５９】

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。

【００６０】

【数１】

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

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

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

【００６１】

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，$y_i \in \{1,-1\}$）は学習データの文脈と分類先を意味し、関数ｓｇｎは、

$$sgn(x) = 1 \quad (x \geqq 0) \qquad (2)$$

$$-1 \quad (otherwise)$$

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。

【００６２】

【数２】

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

$$0 \leq \alpha_i \leq C \ (i = 1,...,l) \qquad (4)$$

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

【００６３】

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を用いる。

【００６４】

$K(x, y) = (x \cdot y + 1)^d$ （６）

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。

【００６５】

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あるから、分類の数が３個以上のデータを扱うために、ペアワイズ手法を組み合わせて用いることになる。本例では、文変換処理システム１５０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た処理を行う。具体的には、ＴｉｎｙＳＶＭを利用して実現する。

［参考文献４：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)］

ステップＳ３：　その後、解を求めたいデータとして入力文３が素性抽出部１１０に入力される。

【００６６】

ステップＳ４：　素性抽出部１１０により、解－素性対抽出部１０１での処理とほぼ同様の処理により入力文３から素性の集合を取り出し、取り出した素性の集合を解推定処理部１１１へ渡す。例えば、入力文３が「犬に噛まれた。」である場合に、以下のような素性を抽出し、抽出した素性の集合を解推定処理部１１１へ渡す。

【００６７】

・　推定すべき格にある体言ｎ＝犬、
・　推定すべき格が修飾する用言ｖ＝噛む、
・　体言ｎと用言ｖとの間の変換前格助詞＝に、

ステップＳ５：　解推定処理部１１１により、学習結果データベース１０３に記憶した学習結果をもとに、渡された素性の集合の場合にどのような解４になりやすいかを推定し、推定された解（変換後格助詞）４を出力する。

【００６８】

例えば、事例「犬に噛まれた。⇒が」、「ヘビに噛まれた。⇒が」の事例について前記のような学習結果が学習結果データベース１０３に記憶されていた場合には、解推定処理部１１１は、この学習結果を参照して、受け取った入力文３から抽出された素性の集合を解析して、変換後格助詞に最もなりやすいのは「が」であると推定して、解４＝「が」を出力する。

【００６９】

図５に、第１の実施の形態における文変換処理システムの別の構成例を示す。なお、以降の図において同一の番号が付与された処理手段などの構成要素は、同一の機能を持つものとする。

【００７０】

文変換処理システム１５０は、素性－解対・素性－解候補対抽出部１６１、機械学習部１６２、学習結果データベース１６３、素性－解候補対抽出部１７０、解推定処理部１７１、および解データベース２を備える。

【００７１】

素性－解対・素性－解候補対抽出部１６１は、解データベース２から事例を取り出し、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手段である。

【００７２】

ここで、解候補は、解以外の解の候補を意味する。すなわち、変換後格助詞となる格助詞が「を」、「に」、「が」、「と」、および「で」の５つであると仮定すると、「が」が解である場合には、「を」、「に」、「と」、および「で」の４つの格助詞が解候補となる。また、解と素性の集合との組を正例と、解候補と素性の集合との組を負例とする。

【００７３】

機械学習部１６２は、素性－解対・素性－解候補対抽出部１６１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との組のときに正例である確率または負例である確率を、サポートベクトルマシン法およびこれに類似する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１６３に記憶する手段である。

【００７４】

素性－解候補抽出部１７０は、入力文３から解候補と素性の集合との組を素性－解対・素性－解候補対抽出部１６１と同様の処理により抽出し、解推定処理部１７１へ渡す手段である。

【００７５】

解推定処理部１７１は、学習結果データベース１６３を参照して、素性－解候補抽出部１７０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、正例である確率が最も大きい解候補を解４と推定し、推定された解４を出力する手段である。

【００７６】

図６に、文変換処理システム１５０の処理フローを示す。

【００７７】

ステップＳ１１：　素性－解対・素性－解候補対抽出部１６１により、解データベース２から事例を取り出し、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。素性－解対・素性－解候補対抽出部１６１により抽出される素性の集合は、ステップＳ１の処理（図２参照）で抽出する素性の集合と同様である。

【００７８】

ステップＳ１２：　機械学習部１６２により、抽出し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機械学習法により学習する。この学習結果を学習結果データベース１６３に記憶する。

【００７９】

例えば、事例が「犬に噛まれた。⇒が」であって、素性の集合が、

・　推定すべき格にある体言ｎ＝犬、
・　推定すべき格が修飾する用言ｖ＝噛む、
・　体言ｎと用言ｖとの間の変換前格助詞＝に、

である場合に、解「が」である確率（正例である確率）と、各解候補「を」、「に」、「と」、および「で」のそれぞれである確率（負例である確率）を求める。

【００８０】

ステップＳ１３：　その後、素性－解候補抽出部１７０に、解を求めたい入力文３が入力される。

【００８１】

ステップＳ１４：　素性－解候補抽出部１７０により、入力文３から解候補と素性の集合との組を、素性－解対・素性－解候補対抽出部１６１と同様の処理により取り出し、取り出した解候補と素性の集合との組を解推定処理部１７１へ渡す。

【００８２】

ステップＳ１５：　解推定処理部１７１により、学習結果データベース１６３に記憶された学習結果をもとに、渡された解候補と素性の集合との組の場合に正例である確率または負例である確率を求める。

【００８３】

例えば、入力文が「犬に噛まれた。」である場合に、抽出した素性の集合と解候補「が」、「を」、「に」、「と」、および「で」それぞれについて、正例である確率または負例である確率を求める。

【００８４】

ステップＳ１６：　すべての解候補に対して正例である確率または負例である確率を求め、正例である確率が最も高い解候補を求める解４として推定し、推定された解４を出力す

る。
【００８５】
〔第２の実施の形態〕
第２の実施の形態として、受け身文・使役文から能動文への変換処理において、教師なし学習により格助詞を自動変換する文変換処理システムの処理を説明する。
【００８６】
まず、機械学習法で用いる教師なしデータを説明する。図７（Ａ）に教師なしデータを作成するために与えられる電子化された文を示す。図７（Ａ）の能動文「犬が私を噛んだ。」は、解析対象となる情報すなわち能動文への文変換時の格助詞の変換に関する情報が付与されていないデータである。しかし、図７（Ａ）の文を能動文への文変換の結果と考えると、この能動文へ変換される元の受け身文・使役文で表れるはずの格助詞（変換前格助詞）は不明であるが、推定すべき解すなわち処理結果（能動文）に表れるべき格助詞（変換格助詞）を抽出することができる。
【００８７】
図７（Ｂ）に変換前格助詞と変換後格助詞との関係を表す単文を示す。図７（Ａ）の能動文の変換元の文は、「犬＜？＞私＜？＞噛んだ（噛まれた）。」と表すことができる。元の文に表れるはずの変換前格助詞は与えられていないことから、「＜？＞（不明）」で示す。また、図７（Ａ）の文から抽出した推定すべき解である変換後格助詞は、＜？＞の下に矢印で示す「が」および「を」で示す。図７（Ｂ）に示すように、解析対象となる情報が与えられていない能動文は、変換前格助詞の情報については不明であるが、解（分類先）である変換後格助詞の情報を持つ。そして、図７（Ｂ）に示す文のうち「犬＜？＞噛んだ。」は、以下のような問題構造に変換することができる。
【００８８】
「問題⇒解」＝「犬＜？＞噛んだ。⇒が」
このように、解析対象の情報が付加されていない能動文を機械学習の教師データとして利用できることがわかる。
【００８９】
図７（Ａ）の能動文から生成される教師なしデータは、変換前格助詞の情報を持たないという点で教師ありデータよりも情報が少ない。しかし、受け身文・使役文に比べて能動文の数が多く、かつ手作業によって変換後格助詞の情報をタグ付けするという作業が不要であるため大量の能動文を教師なしデータとして利用することができ、機械学習法で扱う教師信号を増大させるという利点がある。
【００９０】
図８に、第２の実施の形態における文変換処理システムの構成例を示す。文変換処理システム２００は、ＣＰＵおよびメモリからなり、問題表現相当部抽出部２０１、問題表現情報記憶部２０２、意味解析情報記憶部２０３、問題構造変換部２０４、教師なしデータ記憶部２０５、解－素性対抽出部１０１、機械学習部１０２、学習結果データベース１０３、素性抽出部１１０、解推定処理部１１１、および文データベース５を備える。
【００９１】
問題表現相当部抽出部２０１は、本システムでの処理においてどのようなものが問題表現に相当する部分（問題表現相当部）であるかを予め記憶した問題表現情報記憶部２０２を参照して、解析対象となる情報が付与されていないデータ（文）を記憶した文データベース５から文を取り出し、取り出した文から問題表現相当部を抽出する手段である。
【００９２】
ここでは、問題表現情報記憶部２０２は、問題表現相当部として受け身文・使役文から能動文への変換において変更されるべき格助詞（変換後格助詞）を記憶しておく。
【００９３】
問題構造変換部２０４は、抽出された問題表現相当部を変換する必要がある場合に、意味解析のための情報を記憶する意味解析情報記憶部２０３を参照して、問題表現相当部を変換した文を問題とし問題表現相当部から抽出した格助詞を解として「問題－解」の構造に

変換し、この変換した教師なしデータを事例として教師なしデータ記憶部２０５に記憶する手段である。

【００９４】

文変換処理システム２００の解－素性対抽出部１０１、機械学習部１０２、学習結果データベース１０３、素性抽出部１１０、および解推定処理部１１１は、第１の実施の形態において説明した同一番号の処理手段とほぼ同様の処理を行う手段である。なお、解－素性対抽出部１０１は、教師なしデータ記憶部２０５から、教師なしデータである事例を取り出して各事例ごとに解と素性の集合との組を抽出する。

【００９５】

図９に、教師なしデータ生成処理の処理フローを示す。

【００９６】

ステップＳ２１：　文データベース５から、解析対象となる情報が付与されていない自然文の電子データである文（能動文）が問題表現相当部抽出部２０１に入力される。

【００９７】

ステップＳ２２：　問題表現相当部抽出部２０１により、問題表現情報記憶部２０２を参照し、入力された能動文の構造を検出して問題表現相当部を抽出する。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２０２に記憶されている問題表現情報により与えられる。例えば、問題表現情報として「犬＜？＝推定すべき格（変換後格助詞）＞噛む」を記憶しておく。そして、問題表現相当部抽出部２０１は、問題表現情報として記憶している文構造と入力文（能動文）の構造とをマッチングして、一致するものを問題表現相当部とする。例えば入力文が「犬が噛む。」であれば、マッチングの結果、「が」を問題表現相当部として抽出する。

【００９８】

ステップＳ２３：　問題構造変換部２０４により、意味解析情報記憶部２０３を参照して、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（＜？＞）に変換し、結果として得た文を問題とする。例えば、能動文「犬が噛む。」から問題表現相当部として抽出された「が」を解とし、抽出した「が」の部分を問題表現（＜？＞）に変換し、「犬＜？＞噛む。」を問題とする。

ステップＳ２４：　さらに、問題構造変換部２０４により、この問題および解の構成を持つデータを教師なしデータ（事例）として教師なしデータ記憶部２０５に記憶する。

【００９９】

その後、文変換処理システム２００は、第１の実施の形態における処理と同様に処理を行う（図２参照）。すなわち、解－素性対抽出部１０１により、教師なしデータ記憶部２０５から事例を取り出して、事例ごとに解と素性の集合との組を抽出する（ステップＳ１）。

【０１００】

取り出した事例が、「犬＜？＞噛む。」⇒「が」であれば、例えば以下のような素性の集合を抽出する。

【０１０１】

・推定すべき格にある体言ｎ＝犬、
・推定すべき格が修飾する用言ｖ＝噛む、
・体言ｎと用言ｖの間にあった元の格助詞＝？（不明）。

そして、機械学習部１０２は、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格助詞が解となるかを学習する。機械学習部１０２は、上記のような素性の集合の場合には、「解＝が」になりやすいと学習し、その学習結果を学習結果データベース１０３に記憶する（ステップＳ２）。

【０１０２】

また、取り出した事例が、「ヘビ＜？＞噛む。」⇒「が」であれば、以下のような素性の集合を抽出する。

【０１０３】

・推定すべき格にある体言ｎ＝ヘビ、
・推定すべき格が修飾する用言ｖ＝噛む、
・体言ｎと用言ｖの間にあった元の格助詞＝？（不明）。
そして、機械学習部１０２は、上記のような素性の集合の場合にも、「解＝が」になりやすいと学習し、その学習結果を学習結果データベース１０３に記憶する。
【０１０４】
以降、素性抽出部１１０に入力文３が入力されてから解推定処理部１１１で解４が出力されるまでの処理は、第１の実施の形態における処理として図２の処理フローのステップＳ３～ステップＳ５に示す処理と同様であるので説明を省略する。
【０１０５】
図１０に、第２の実施の形態における文変換処理システムの別の構成例を示す。文変換処理システム２５０は、問題表現相当部抽出部２０１、問題表現情報記憶部２０２、意味解析情報記憶部２０３、問題構造変換部２０４、教師なしデータ記憶部２０５、素性－解対・素性－解候補対抽出部１６１、機械学習部１６２、学習結果データベース１６３、素性－解候補抽出部１７０、解推定処理部１７１、および文データベース５を備える。
【０１０６】
文変換処理システム２５０の問題表現相当部抽出部２０１、問題表現情報記憶部２０２、意味解析情報記憶部２０３、および問題構造変換部２０４は、図８に示す同一の番号が付与された各処理手段と同様の処理を行う手段である。
【０１０７】
また、文変換処理システム２５０の素性－解対・素性－解候補対抽出部１６１、機械学習部１６２、学習結果データベース１６３、素性－解候補抽出部１７０、および解推定処理部１７１は、図５に示す同一の番号が付与された各処理手段とほぼ同様の処理を行う手段である。
【０１０８】
文変換処理システム２５０は、素性－解対・素性－解候補対抽出部１６１により、教師なしデータ記憶部２０５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（図６：ステップＳ１１）。
【０１０９】
取り出した事例が、「犬＜？＞噛む。」⇒「が」であれば、例えば以下のような素性の集合を抽出する。
【０１１０】
・推定すべき格にある体言ｎ＝犬、
・推定すべき格が修飾する用言ｖ＝噛む、
・体言ｎと用言ｖの間にあった元の格助詞＝？（不明）。
そして、機械学習部１６２により、解もしくは解候補と素性の集合の組から、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機械学習法により学習する。この学習結果を学習結果データベース１６３に記憶する（図６：ステップＳ１２）。
【０１１１】
以降、素性－解候補抽出部１７０に入力文３が入力されてから解推定処理部１７１で解４が出力されるまでの処理は、第１の実施の形態における処理として図６の処理フローのステップＳ１３～ステップＳ１６の処理と同様であるので説明を省略する。
【０１１２】
〔第３の実施の形態〕
教師なしデータ記憶部２０５に記憶される事例（「問題－解」）は、解データベース２に記憶されている事例（「問題－解」）とほとんど同じ構造であることから、教師なしデータの事例と教師ありデータの事例とを混ぜ合わせて利用することも可能である。本形態で、教師なしデータおよび教師ありデータの両方を教師信号として用いて機械学習を行う方法を、「教師あり／なし学習」と呼ぶ。

【０１１３】
教師なしデータは、元の文に表れる変換前格助詞の情報を持たず、教師ありデータよりも情報が少ない。しかし、人手により事例ごとに解情報（変換後格助詞など）をタグ付けする必要がない。また、一般的に受け身文の数より能動文の数が多いため、多くの文を教師信号として利用できる。このため、教師あり／なし学習による文変換処理は、人手により解析対象の情報を付与するという労力負担を増やすことなく大量の教師データを用いた機械学習の学習結果を用いた文変換処理を行うことができるという利点がある。
【０１１４】
図１１に、第３の実施の形態における文変換処理システム３００の構成例を示す。文変換処理システム３００は、ＣＰＵおよびメモリからなり、問題表現相当部抽出部２０１、問題表現情報記憶部２０２、意味解析情報記憶部２０３、問題構造変換部２０４、教師なしデータ記憶部２０５、解－素性対抽出部１０１、機械学習部１０２、学習結果データベース１０３、素性抽出部１１０、解推定処理部１１１、解データベース２、および文データベース５を備える。文変換処理システム３００は、第２の実施の形態として説明した図８に示す構成を備える文変換処理システム２００に、さらに解データベース２を備えた構成をとり、文変換処理システム２００とほぼ同様の処理を行う。
【０１１５】
解－素性対抽出部１０１は、解データベース２に記憶された教師ありデータである事例および教師なしデータ記憶部２０５に記憶された教師なしデータである事例について、事例ごとに解と素性の集合との組を抽出する。
【０１１６】
図１２に、第３の実施の形態における文変換処理システムの別の構成例を示す。文変換処理システム３５０は、ＣＰＵおよびメモリからなり、問題表現相当部抽出部２０１、問題表現情報記憶部２０２、意味解析情報記憶部２０３、問題構造変換部２０４、教師なしデータ記憶部２０５、素性－解対・素性－解候補対抽出部１６１、機械学習部１６２、学習結果データベース１６３、素性－解候補抽出部１７０、解推定処理部１７１、解データベース２、および文データベース５を備える。
【０１１７】
文変換処理システム３５０は、第２の実施の形態として説明した図１０に示す構成を備える文変換処理システム２５０に、さらに解データベース２を備えた構成をとり、文変換処理システム２５０とほぼ同様の処理を行う。
【０１１８】
素性－解対・素性－解候補対抽出部１６１は、解データベース２に記憶された教師ありデータである事例および教師なしデータ記憶部２０５に記憶された教師なしデータである事例について、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】
〔第４の実施の形態〕
第４の実施の形態として、言語解析処理を行う場合に、教師なしデータおよび教師ありデータの両方の利点を活かしたスタック型機械学習を行って解析処理を行う言語解析処理システムの処理を説明する。
【０１２０】
スタック型機械学習は、複数のシステムの解析結果の融合に用いられている「スタッキング」と呼ばれる手法を用いた機械学習であって、異なる機械学習法の解析結果を素性に追加した教師信号を用いて機械学習を行うものである。
［参考文献５：Hans van Halteren, Jakub, Zavrel, and Walter Daelemans, Improving Accuracy in Word Class Tagging Through the Combination of Machine Learning Systems, Computational Linguistics, Vol.27, No.2, (2001), pp.199-229 ］
本形態において、言語解析処理システムは、借用型機械学習（教師なしデータを用いた機械学習）または併用型機械学習（教師あり／なしデータによる機械学習）を用いた言語解析処理を行い、その処理結果である推定解を素性の集合の要素として追加する。そして、

推定解が追加された素性の集合を用いてさらに教師あり学習による言語解析処理を行う。

【０１２１】

例えば、本形態の言語解析処理システムで用いられる教師あり機械学習において、ある教師ありデータ（事例）から抽出される素性の集合がリスト｛ａ，ｂ，ｃ｝を持つとする。そして、スタッキング用処理システムが教師なし機械学習を用いた言語解析処理システムであり、その解析結果が「d_1」であるとする。この場合に、言語解析処理システムの教師あり機械学習処理では、素性の集合｛ａ，ｂ，ｃ｝に解析結果「d_1」を追加し、リスト｛ａ，ｂ，ｃ，”教師なし学習の解析結果＝d_1”｝を新しい素性の集合として機械学習を行なう。

【０１２２】

また、スタッキング用処理システムが教師あり／なし機械学習を用いた言語解析処理システムであり、その解析結果が「d_2」であるとする。この場合に、言語解析処理システムの教師あり機械学習処理では、素性の集合｛ａ，ｂ，ｃ｝に解析結果「d_2」を追加し、リスト｛ａ，ｂ，ｃ，”教師あり／なし学習の解析結果＝d_2”｝を新しい素性の集合として機械学習を行なう。

【０１２３】

また、スタッキング用処理システムとして、教師なし機械学習を用いた言語解析処理システムと、教師あり／なし機械学習を用いた言語解析処理システムとを利用することも可能である。この場合に、言語解析処理システムの教師あり機械学習処理では、素性の集合｛ａ，ｂ，ｃ｝に解析結果「d_1」および「d_2」を追加し、リスト｛ａ，ｂ，ｃ，”教師なし学習の解析結果＝d_1”，”教師あり／なし学習の解析結果＝d_2”｝を新しい素性の集合として機械学習を行なう。

【０１２４】

このように、スタッキング手法を用いて、教師ありデータを用いた非借用型機械学習と借用型機械学習または併用型機械学習とを組み合わせた場合には、教師あり機械学習に用いる教師ありデータ（事例）の素性が増加する。これにより、教師あり機械学習に用いる個々の事例自体が学習精度を向上させると考えられる。さらに、教師あり機械学習では、素性が増加してはいるが教師ありデータ（事例）についての正解率を最大にするような学習、すなわち解析処理対象についての精度を最大にするような学習を行い、その学習結果を用いて解析処理を行う。これにより、教師あり機械学習、教師なし機械学習それぞれの利点をうまく利用して高い解析精度を得ることが期待できる。

【０１２５】

図１３に、第４の実施の形態における言語解析処理システムの構成例を示す。

【０１２６】

言語解析処理システム５００は、与えられた問題に対する言語解析処理の解析結果を出力するシステムであって、ＣＰＵおよびメモリからなり、解－素性対抽出部５０１、機械学習部５０２、学習結果データベース５０３、素性抽出部５０４、解推定処理部５０５、スタック用教師なし学習処理システム１０１０、第１素性追加部５１１、第２素性追加部５１２、文データベース５、および解データベース６を備える。

【０１２７】

解－素性対抽出部５０１、機械学習部５０２、学習結果データベース５０３、素性抽出部５０４、および解推定処理部５０５の各処理手段は、それぞれ、文変換処理システム１００の解－素性対抽出部１０１、機械学習部１０２、学習結果データベース１０３、素性抽出部１１０、および解推定処理部１１１とほぼ同様の処理を行う手段である。

【０１２８】

スタック用教師なし学習処理システム１０１０は、言語解析処理について、文データベース５から生成した教師なしデータから素性の集合を抽出し、抽出された素性の集合からどのような素性の集合のときにどのような解（解析結果）になりやすいかを学習してその学習結果を記憶しておき、第１素性追加部５１１または第２素性追加部５１２から受け取った素性の集合の場合にどのような解（解析結果）になりやすいかを記憶しておいた学習結

果から推定し、推定された解d_1を第１素性追加部５１１へまたは解d_1’を第２素性追加部５１２へ返却する手段である。

【０１２９】

スタック用教師なし学習処理システム１０１０は、図８に示す文変換処理システム２００と同様に構成された処理手段、すなわち問題表現相当部抽出部２０１、問題表現情報記憶部２０２、意味解析情報記憶部２０３、問題構造変換部２０４、教師なしデータ記憶部２０５、解－素性対抽出部１０１、機械学習部１０２、学習結果データベース１０３、素性抽出部１１０、および解推定処理部１１１を備え（図示しない）、与えられた問題に対する言語解析処理の解析結果を出力する。

【０１３０】

第１素性追加部５１１は、解－素性対抽出部５０１から受け取った解と素性の集合の組から素性の集合のみを取り出してスタック用教師なし学習処理システム１０１０へ渡し、スタック用教師なし学習処理システム１０１０から返却された解d_1を受け取り、”教師なし学習の解析結果＝d_1”を素性として元の素性の集合に追加する手段である。

【０１３１】

第２素性追加部５１２は、素性抽出部５０４から受け取った素性の集合を取り出してスタック用教師なし学習処理システム１０１０へ渡し、スタック用教師なし学習処理システム１０１０から返却された解d_1’を受け取り、”教師なし学習の解析結果＝d_1’”を素性として素性の集合に追加する手段である。

【０１３２】

図１４および図１５に、言語解析処理システム５００の処理フローを示す。

【０１３３】

ステップＳ３０：　スタック用教師なし学習処理システム１０１０では、文データベース５に格納された単文を取り出す。取り出した文から問題表現情報を参照して問題表現相当部を抽出して解とし、意味解析情報を参照して問題表現相当部を問題構造に変換して結果として得た文を問題とし、この「問題－解」構造を持つ事例を教師なしデータとして記憶する。さらに、各事例ごとに解と素性の集合との組を抽出し、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、学習結果を記憶しておく。

【０１３４】

ステップＳ３１：　その後、解－素性対抽出部５０１により、解データベース６から事例を取り出し、各事例ごとに解と素性の集合との組を抽出する。

【０１３５】

ステップＳ３２：　第１素性追加部５１１により、解と素性の集合との組のうち素性の集合のみを取り出し、スタック用教師なし学習処理システム１０１０へ渡す。

【０１３６】

ステップＳ３３：　スタック用教師なし学習処理システム１０１０では、予め記憶しておいた学習結果を参照して、受け取った素性の集合についてどのような解になりやすいかを推定し、推定された解d_1を第１素性追加部５１１へ返却する。

【０１３７】

ステップＳ３４：　第１素性追加部５１１により、返却された解d_1を素性として元の素性の集合に追加する。その結果、元の素性の集合が｛ａ，ｂ，ｃ｝であるとすると、機械学習部５０２に渡される素性の集合は、｛ａ，ｂ，ｃ，”教師なし学習の解析結果＝d_1”｝となる。

【０１３８】

ステップＳ３５：　機械学習部５０２により、解と”教師なし学習の解析結果＝d_1”を含む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学習し、学習結果を学習結果データベース５０３に記憶する。

【０１３９】

ステップＳ３６：　解を求めたい文が素性抽出部５０４に入力される。

【０１４０】

ステップＳ３７：　素性抽出部５０４により、入力文３から素性の集合を取り出して、第２素性追加部５１２へ渡す。

【０１４１】

ステップＳ３８：　第２素性追加部５１２により、受け取った素性の集合がスタック用教師なし学習処理システム１０１０へ渡される。

【０１４２】

ステップＳ３９：　スタック用教師なし学習処理システム１０１０では、予め記憶しておいた学習結果を参照して、受け取った素性の集合のときにどのような解となりやすいかを推定し、推定された解d_1’を第２素性追加部５１２へ渡す。

【０１４３】

ステップＳ３１０：　第２素性追加部５１２により、返却された解d_1’を素性として元の素性の集合に追加する。元の素性の集合が｛ａ，ｂ，ｃ｝であるとすると、機械学習部５０２に渡される素性の集合は、｛ａ，ｂ，ｃ，”教師なし学習の解析結果＝d_1’”｝となり、この素性の集合が解推定処理部５０５へ渡される。

【０１４４】

ステップＳ３１１：　解推定処理部５０５により、学習結果データベース５０３に記憶された学習結果を参照して、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、推定された解４を出力する。

【０１４５】

以下に、具体的な処理を例として言語解析処理システム５００の処理をより詳細に説明する。第１の具体例として、言語解析処理システム５００が受け身文・使役文から能動文への変換処理における変換後格助詞の推定を行う場合の処理例を示す。

【０１４６】

言語解析処理システム５００のスタック用教師なし学習処理システム１０１０では、予め受け身文・使役文から能動文への変換処理において変換すべき格助詞（推定すべき格助詞）を問題表現として記憶しておく。そして、文データベース５から取り出した文が「犬が噛む」であるときには、問題表現相当部として「が」を抽出して解（分類先）とし、文を「犬＜？＞噛む。」に変形して問題（文脈）とし、

事例（問題⇒解）：「犬＜？＞噛む。」⇒「が」

を記憶する。さらに、この事例から以下のような素性の集合を抽出する。

【０１４７】

・推定すべき格にある体言ｎ＝犬、

・推定すべき格が修飾する用言ｖ＝噛む、

・体言と用言の間の元の（変換前）格助詞＝？（不明）

そして、この素性の集合の場合には変換後格助詞は「が」になりやすいと学習し、その学習結果を記憶する。

【０１４８】

また、文データベース５から取り出した文が「ヘビが噛む」であるときには、同様の処理により、

事例（問題⇒解）：「ヘビ＜？＞噛む。」⇒「が」

を記憶する。さらに、この事例から、以下のような素性の集合を抽出する。

【０１４９】

・推定すべき格にある体言ｎ＝ヘビ、

・推定すべき格が修飾する用言ｖ＝噛む、

・体言と用言の間の元の（変換前）格助詞＝？（不明）

そして、この素性の集合の場合にも変換後格助詞は「が」になりやすいと学習し、その学習結果を記憶する。

【０１５０】

その後、解－素性対抽出部５０１により、解データベース６から、

事例（問題⇒解）：「犬に噛まれる。」⇒「が」

を取り出し、各事例ごとに解「が」と以下の素性の集合との組を抽出する。

【０１５１】

・推定すべき格にある体言ｎ＝犬、
・推定すべき格が修飾する用言ｖ＝噛む、
・体言ｎと用言ｖとの間の元の（変換前）格助詞＝に

さらに、第１素性追加部５１１により、抽出した解と素性の集合との組のうち、素性の集合のみを取り出し、スタック用教師なし学習処理システム１０１０へ渡す。スタック用教師なし学習処理システム１０１０では、予め記憶しておいた学習結果を参照して、受け取った素性の集合についてどのような解になりやすいかを推定し、推定された解d_1「が」を第１素性追加部５１１へ返却する。

【０１５２】

次に、第１素性追加部５１１により、返却された解d_1を素性として元の素性の集合に追加し、以下のような素性の集合とする。

【０１５３】

・推定すべき格にある体言ｎ＝犬、
・推定すべき格が修飾する用言ｖ＝噛む、
・体言ｎと用言ｖとの間の元の（変換前）格助詞＝に、
・教師なし学習の解析結果＝が（解d_1）

そして、機械学習部５０２により、解と解d_1を含む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学習し、学習結果を学習結果データベース５０３に記憶する。

【０１５４】

その後、解を求めたい文が素性抽出部５０４に入力される。素性抽出部５０４により、入力文３から素性の集合を取り出す。例えば、入力文３が「ヘビに噛まれる。」である場合に、以下のような素性の集合を抽出して、第２素性追加部５１２へ渡す。

【０１５５】

・推定すべき格にある体言ｎ＝ヘビ、
・推定すべき格が修飾する用言ｖ＝噛む、
・体言ｎと用言ｖとの間の元の（変換前）格助詞＝に

そして、第２素性追加部５１２により、受け取った素性の集合がスタック用教師なし学習処理システム１０１０へ渡される。スタック用教師なし学習処理システム１０１０では、予め記憶しておいた学習結果を参照して、受け取った素性の集合のときにどのような解となりやすいかを推定し、推定された解d_1’「が」を第２素性追加部５１２へ返却する。

【０１５６】

第２素性追加部５１２により、返却された解d_1’を素性として元の素性の集合に追加する。例えば、以下のような素性の集合となる。

【０１５７】

・推定すべき格にある体言ｎ＝ヘビ、
・推定すべき格が修飾する用言ｖ＝噛む、
・体言ｎと用言ｖとの間の元の（変換前）格助詞＝に、
・教師なし学習の解析結果＝が（解d_1’）

そして、解d_1’を含む素性の集合は、解推定処理部５０５へ渡される。解推定処理部５０５により、学習結果データベース５０３に記憶された学習結果を参照して、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て、推定された解４を出力する。

【０１５８】

ここでは、スタック用教師なし学習処理システム１０１０から返却された解析結果「が」を追加した素性の集合をもとに教師あり学習の学習結果を参照して推定した格助詞「が」が出力される。

【０１５９】

このように、機械学習部５０２は、解データベース６の教師ありデータ（事例）から抽出

した素性の集合に”教師なし学習の解析結果＝d_1”を追加した素性の集合を用いて機械学習を行う。この場合に用いる素性の集合は、教師ありデータから抽出した素性の集合よりも素性の情報が多くなるため、教師ありデータのみを用いて機械学習を行う場合に比べてより高い精度で機械学習を行うことができる。また、データ量は膨大であるが素性の情報が少ない教師なしデータのみを用いて機械学習を行う場合に比べても、素性の情報が多い点でより高い精度の機械学習を行うことができる。

【０１６０】

さらに、解推定処理部５０５は、素性の集合の情報が多い事例を用いて学習された高い精度の学習結果を参照して、入力文３から抽出した素性の集合の類似性をみることになる。したがって、素性の集合に”教師なし学習の解析結果＝d_1’”を含まない場合に比べて、素性の集合同士の類似性が高くなり、推定処理の精度も高くなる。

【０１６１】

第２の具体例として、言語解析処理システム５００が、文の意味が深層格などで表現されている場合に、その文を生成する際に与えられる表層格を推定する処理を行う場合の処理例を示す。

【０１６２】

例えば、文の意味を深層格で示すと以下のように表すことができる。

【０１６３】

文「りんご＜←ｏｂｊ＞食べる」

この文において、「りんご」は「食べる」の目的語であり、「りんご」と「食べる」とは深層格の目的格（＜←ｏｂｊ＞で示す。）で連結されている。

【０１６４】

そして、文生成処理では、前記の元の文から、生成文「りんごを食べる」を出力するが、この場合に＜←ｏｂｊ＞に対応する格助詞「を」を生成する必要がある。この処理において与えられる問題構造（問題⇒格）を以下に示す。

【０１６５】

問題（問題⇒格）：

「りんご＜←ｏｂｊ＞食べる」⇒「を」

言語解析処理システム５００のスタック用教師なし学習処理システム１０１０は、与えられている深層格を問題表現として記憶しておく。そして、スタック用教師なし学習処理システム１０１０では、文データベース５から取り出した文が「りんごを食べる。」である場合に、格助詞「を」を問題表現相当部として置き換え、格助詞「を」を解として抽出し、取り出した文の問題表現相当部を変換した結果得た文を問題として、以下のような事例を教師なしデータとして記憶する。

【０１６６】

事例（問題⇒解）：

「りんご＜？＞食べる」⇒「を」

さらに、この事例から解と素性の集合との組を抽出する。ここで、素性の集合は、以下のようになる。

【０１６７】

・生成すべき格にある体言ｎ＝りんご、

・生成すべき格が修飾する用言ｖ＝食べる、

・体言ｎと用言ｖの間の深層格＝？（不明）

そして、どのような素性の集合のときにどのような解となりやすいかを学習し、その学習結果を記憶しておく。例えば、前記の素性の集合の場合には「解＝を」になりやすいと学習する。

【０１６８】

また、文データベース５から文「みかんを食べる」を取り出したとする。この場合には、以下のような事例を教師なしデータとする。

【０１６９】

事例（問題⇒解）：
「みかん＜？＞食べる」⇒「を」
さらに、この事例から解と素性の集合との組を抽出する。ここで、素性の集合は、以下のようになる。
【０１７０】
・生成すべき格にある体言ｎ＝みかん、
・生成すべき格が修飾する用言ｖ＝食べる、
・体言ｎと用言ｖの間の深層格＝？（不明）
なお、文生成処理における格推定の場合にも、一般的な教師ありデータに比べて素性の情報は少なくなるが、教師なしデータとして利用できる文自体は多量にあるため、多数の教師なしデータを準備することが可能である。
【０１７１】
そして、どのような素性の集合のときにどのような解となりやすいかを学習し、その学習結果を記憶しておく。この場合にも、「解＝を」になりやすいと学習する。
【０１７２】
その後、解－素性対抽出部５０１により、解データベース６から以下の事例を取り出したとする。
【０１７３】
事例：「りんご＜←ｏｂｊ＞食べる」⇒「を」
さらに、取り出した事例から解と素性の集合との組を抽出する。素性の集合として以下のものが抽出される。
【０１７４】
・生成すべき格にある体言ｎ＝りんご、
・生成すべき格が修飾する用言ｖ＝食べる、
・体言ｎと用言ｖの間の深層格＝ｏｂｊ
第１素性追加部５１１により、抽出した素性の集合をスタック用教師なし学習処理システム１０１０へ渡し、スタック用教師なし学習処理システム１０１０では、記憶しておいた学習結果をもとに、受け取っ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、推定された解d_1＝「を」を第１素性追加部５１１へ返却する。そして、第１素性追加部５１１は、返却された解d_1を素性の集合に追加して、以下の素性の集合とする。
【０１７５】
・生成すべき格にある体言ｎ＝りんご、
・生成すべき格が修飾する用言ｖ＝食べる、
・体言ｎと用言ｖの間の深層格＝ｏｂｊ、
・教師なし学習の解析結果＝を（解d_1）
そして、機械学習部５０２は、前記の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する。このとき、スタック用教師なし学習処理システム１０１０から取得した解d_1による”教師なし学習の解析結果＝を（解d_1）”を素性の集合として持つため、
・生成すべき格にある体言ｎ＝りんご、
・生成すべき格が修飾する用言ｖ＝食べる、
・体言ｎと用言ｖの間の深層格＝ｏｂｊ、
・教師なし学習の解析結果＝を（解d_1）
という素性があれば、「を」が解となるという学習ができている。この学習結果を学習結果データベース５０３に記憶する。
【０１７６】
その後、素性抽出部５０４に文「みかん＜←ｏｂｊ＞食べる」が入力されると、素性抽出部５０４は、入力文３から、以下のような素性の集合を抽出して、第２素性追加部５１２へ渡す。
【０１７７】
・生成すべき格にある体言ｎ＝みかん、

・生成すべき格が修飾する用言ｖ＝食べる、
・体言ｎと用言ｖの間の深層格＝ｏｂｊ
第２素性追加部５１２により、この素性の集合がスタック用教師なし学習処理システム１０１０に渡されると、スタック用教師なし学習処理システム１０１０では、記憶しておいた学習結果を参照して受け取った素性の集合の場合になりやすい解d_1’＝「を」を推定し、第２素性追加部５１２へ返却する。
【０１７８】
第２素性追加部５１２は、元の素性の集合に解d_1’を追加した以下の素性の集合を解推定処理部５０５へ渡す。
【０１７９】
・生成すべき格にある体言ｎ＝みかん、
・生成すべき格が修飾する用言ｖ＝食べる、
・体言ｎと用言ｖの間の深層格＝ｏｂｊ、
・教師なし学習の解析結果＝を（解d_1’）
解推定処理部５０５により、この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する。ここで、学習結果として記憶しておいた素性の集合と、入力文３から抽出した素性の集合とがよく類似しているので、学習結果で解とした「を」を正しく推定することができる。そして、推定された解４として生成すべき格助詞「を」を出力する。
【０１８０】
次に、第３の具体例として、言語解析処理システム５００が、動詞の省略表現を補完する処理を行う場合の処理例を示す。例えば、「そんなにうまくいくとは。」という文は文末の動詞部分が省略されている表現であると考えて、省略された動詞部分「思えない」を補完する処理を行う。
【０１８１】
この場合に、省略された「補完すべき動詞部分」を問題表現とし、その省略表現を補完する「動詞部分」を解とする。言語解析処理システム５００のスタック用教師なし学習処理システム１０１０では、このような問題表現を抽出するために予め問題表現情報を記憶しておく。
【０１８２】
そして、文データベース５から取り出した文が「そんなにうまくいくとは思えない。」である場合に、文末の動詞部分を問題表現相当部として置き換え、文末の動詞部分「思えない」を解として抽出し、取り出した文の問題表現相当部を変換した結果得た文を問題として、以下のような事例を教師なしデータとして記憶する。
【０１８３】
事例（問題⇒解）：
「そんなにうまくいくとは＜？＞」⇒「思えない」
さらに、この事例から解と素性の集合との組を抽出する。ここで、素性の集合は、以下のようになる。
【０１８４】
・「は」、
・「とは」、
・「くとは」、
・「いくとは」、
…、
・「そんなにうまくいくとは思えない」
そして、どのような素性の集合のときにどのような解となりやすいかを学習し、その学習結果を記憶しておく。例えば、前記の素性の集合の場合には「解＝思えない」になりやすいと学習する。
【０１８５】
その後、解－素性対抽出部５０１により、解データベース６から、

事例：「そんなにうまくいくとは。」⇒「思えない」
を取り出し、取り出した事例から解と素性の集合との組を抽出する。ここで、素性の集合は、以下の素性からなる。
【０１８６】
・「は」、
・「とは」、
・「くとは」、
・「いくとは」、
…、
・「そんなにうまくいくとは」
・「そんなにうまくいくとは思えない」
第１素性追加部５１１は、抽出した素性の集合をスタック用教師なし学習処理システム１０１０へ渡す。
【０１８７】
スタック用教師なし学習処理システム１０１０では、記憶しておいた学習結果をもとに、受け取っ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、推定された解d_1＝「思えない」を第１素性追加部５１１へ返却する。
【０１８８】
そして、第１素性追加部５１１は、返却された解d_1を素性の集合に追加して、以下の素性の集合とする。
【０１８９】
・「は」、
・「とは」、
・「くとは」、
・「いくとは」、
…、
・「そんなにうまくいくとは」
・「そんなにうまくいくとは思えない」
・教師なし学習の解析結果＝思えない（解d_1）
そして、機械学習部５０２は、前記の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し、学習結果を学習結果データベース５０３に記憶する。
【０１９０】
その後、素性抽出部５０４に文「そううまくいくとは。」が入力されると、素性抽出部５０４は、入力文３から、以下のような素性の集合を抽出して、第２素性追加部５１２へ渡す。
【０１９１】
・「は」、
・「とは」、
・「くとは」、
・「いくとは」、
…、
・「そううまくいくとは」
第２素性追加部５１２により、この素性の集合がスタック用教師なし学習処理システム１０１０に渡されると、スタック用教師なし学習処理システム１０１０では、記憶しておいた学習結果を参照して受け取った素性の集合の場合になりやすい解d_1’＝「思えない」を推定し、第２素性追加部５１２へ返却する。
【０１９２】
第２素性追加部５１２は、元の素性の集合に解d_1’を追加した以下の素性の集合を解推定処理部５０５へ渡す。
【０１９３】

・「は」、
・「とは」、
・「くとは」、
・「いくとは」、
…、
・「そううまくいくとは」
・教師なし学習の解析結果＝思えない（解d_1’）
解推定処理部５０５により、この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、推定された解４として省略された動詞部分「思えない」を出力する。

【０１９４】
図１６に、第４の実施の形態における言語解析処理システムの別の構成例を示す。言語解析処理システム５４０は、言語解析処理システム５００と同様の処理手段を備え、スタック用教師なし学習処理システム１０１０の代わりに、スタック用教師あり／なし学習処理システム１０２０を備えた構成をとる。

【０１９５】
スタック用教師あり／なし学習処理システム１０２０は、スタック用教師なし学習処理システム１０１０と同様の処理手段に解データベース２を追加した構成をとる。スタック用教師あり／なし学習処理システム１０２０は、言語解析処理について、文データベース５から生成した教師なしデータおよび解データベース２の事例（教師ありデータ）からそれぞれ素性の集合を抽出し、抽出された素性からどのような素性の集合のときにどのような解（解析結果）になりやすいかを学習してその学習結果を記憶しておき、第１素性追加部５１１または第２素性追加部５１２から受け取った素性の集合の場合にどのような解（解析結果）になりやすいかを記憶しておいた学習結果から推定し、推定された解d_2を第１素性追加部５１１へ、または解d_2’を第２素性追加部５１２へ返却する手段である。

【０１９６】
言語解析処理システム５４０の第１素性追加部５１１は、スタック用教師あり／なし学習処理システム１０２０から返却された解d_2を受け取り、”教師あり／なし学習の解析結果＝d_2”を素性として元の素性の集合に追加する。また、言語解析処理システム５４０の第２素性追加部５１２は、スタック用教師あり／なし学習処理システム１０２０から返却された解d_2’を受け取り、”教師あり／なし学習の解析結果＝d_2’”を素性として素性の集合に追加する。

【０１９７】
さらに、図１７に、第４の実施の形態における言語解析処理システムの別の構成例を示す。

【０１９８】
言語解析処理システム５５０は、与えられた問題に対する言語解析処理の解析結果を出力システムであって、ＣＰＵおよびメモリからなり、素性－解対・素性－解候補対抽出部５６１、機械学習部５６２、学習結果データベース５６３、素性－解候補抽出部５６４、解推定処理部５６５、スタック用教師なし学習処理システム１０３０、第１素性追加部５２１、第２素性追加部５２２、文データベース５、および解データベース６を備える。

【０１９９】
素性－解対・素性－解候補対抽出部５６１、機械学習部５６２、学習結果データベース５６３、素性－解候補抽出部５６４、および解推定処理部５６５の各処理手段は、それぞれ、文変換処理システム１５０の素性－解対・素性－解候補対抽出部１６１、機械学習部１６２、学習結果データベース１６３、素性－解候補抽出部１７０、および解推定処理部１７１とほぼ同様の処理を行う手段である。

【０２００】
スタック用教師なし学習処理システム１０３０は、言語解析処理について、文データベース５から生成した教師なしデータから解もしくは解候補と素性の集合との組を抽出し、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機械学習法により学習してその学習結果を記憶しておき、この学習結果を参照して第１素性追加部５２１または第２素性追加部５２２から受け取った解もしくは解候補と素性の集合との組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（解析結果）と推定し、推定された解d_3を第１素性追加部５２１へまたは解d_3’を第２素性追加部５２２へ返却する手段である。

【０２０１】

スタック用教師なし学習処理システム１０３０は、解d_3、解d_3’として、解と推定した解候補を出力するとともに、その解が正例もしくは負例であるかの情報や、正例もしくは負例である確率の情報などを出力することもできる。

【０２０２】

スタック用教師なし学習処理システム１０３０は、図１０に示す文変換処理システム２５０と同様に構成された処理手段、すなわち問題表現相当部抽出部２０１、問題表現情報記憶部２０２、意味解析情報記憶部２０３、問題構造変換部２０４、教師なしデータ記憶部２０５、素性－解対・素性－解候補対抽出部１６１、機械学習部１６２、学習結果データベース１６３、素性－解候補抽出部１７０、および解推定処理部１７１を備え（図示しない）、与えられた問題に対する言語解析処理の解析結果を出力する。

【０２０３】

第１素性追加部５２１は、素性－解対・素性－解候補対抽出部５６１から受け取った解もしくは解候補と素性の集合との組をスタック用教師なし学習処理システム１０３０へ渡し、スタック用教師なし学習処理システム１０３０から返却された解d_3を受け取り、”教師なし学習の解析結果＝解d_3”を素性として元の素性の集合に追加する手段である。

【０２０４】

第２素性追加部５２２は、素性－解候補抽出部５６４から受け取った解候補と素性の集合との組をスタック用教師なし学習処理システム１０３０へ渡し、スタック用教師なし学習処理システム１０３０から返却された解d_3’を受け取り、”教師なし学習の解析結果＝解d_3’”を素性として元の素性の集合に追加する手段である。

【０２０５】

図１８および図１９に、言語解析処理システム５５０の処理フローを示す。

【０２０６】

ステップＳ４０：　スタック用教師なし学習処理システム１０３０では、文データベース５に格納された単文を取り出し、取り出した文から問題表現情報を参照して問題表現相当部を抽出して解とし、さらに意味解析情報を参照して問題表現相当部を問題構造に変換し、変換結果として得た文を問題として「問題－解」構造を持つ事例を教師なしデータとして記憶する。さらに、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し、どのような解もしくは解候補と素性の集合との組のときに正例である確率または負例である確率を機械学習法により学習し、学習結果を記憶しておく。

【０２０７】

ステップＳ４１：　その後、素性－解対・素性－解候補対抽出部５６１により、解データベース６から事例を取り出し、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。

【０２０８】

ステップＳ４２：　第１素性追加部５２１により、解もしくは解候補と素性の集合との組をスタック用教師なし学習処理システム１０３０へ渡す。

【０２０９】

ステップＳ４３：　スタック用教師なし学習処理システム１０３０では、予め記憶しておいた学習結果を参照して、受け取った解もしくは解候補と素性の集合との組について正例である確率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d_3と推定し、解d_3を第１素性追加部５２１へ返却する。

【０２１０】

ステップＳ４４：　第１素性追加部５２１により、返却された解d_3から、”教師なし学習の解析結果＝解d_3”を素性として元の素性の集合に追加する。解d_3として、推定された解候補の他に、正例もしくは負例であるかの情報、正例もしくは負例である確率などの情報が含まれている場合には、受け取った解d_3に含まれる情報の一部または全部を素性の集合に追加するようにしてもよい。例えば、”教師なし学習の解析結果＝推定された解候補（解d_3）”、”教師なし学習の解析結果＝正例／負例（解d_3）”、または”教師なし学習の解析結果＝正例の確率／負例の確率（解d_3）”のような素性の１つもしくは複数が元の素性の集合に追加される。

【０２１１】

ステップＳ４１～ステップＳ４４の処理は、すべての解もしくは解候補と素性の集合との組について行なわれる。

【０２１２】

ステップＳ４５：　機械学習部５６２により、解もしくは解候補と解d_3を含む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組のときに正例である確率または負例である確率を機械学習法により求め、その学習結果を学習結果データベース５６３に記憶する。

【０２１３】

ステップＳ４６：　解を求めたい文が素性－解候補抽出部５６４に入力される。

【０２１４】

ステップＳ４７：　素性－解候補抽出部５６４により、入力文３から解候補と素性の集合との組を取り出す。

【０２１５】

ステップＳ４８：　第２素性追加部５２２により、受け取った解候補と素性の集合との組をスタック用教師なし学習処理システム１０３０へ渡す。

【０２１６】

ステップＳ４９：　スタック用教師なし学習処理システム１０３０では、予め記憶しておいた学習結果を参照して、受け取った解候補と素性の集合との組からどのような解候補と素性の集合との組のときに正例である確率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d_3’と推定し、解d_3’を第２素性追加部５２２へ返却する。

【０２１７】

ステップＳ４１０：　第２素性追加部５２２により、返却された解d_3’から、”教師なし学習の解析結果＝解d_3’”を素性として元の素性の集合に追加する。

【０２１８】

ステップＳ４１１：　解推定処理部５６５により、学習結果データベース５６３に記憶された学習結果を参照して、渡された解候補と素性の集合との場合に正例である確率または負例である確率を求める。すべての解候補についてこの確率を求め、正例である確率が最も大きい解候補を求める解４として出力する。

【０２１９】

図２０に、第４の実施の形態における言語解析処理システムの別の構成例を示す。言語解析処理システム５８０は、言語解析処理システム５５０と同様の処理手段を備え、スタック用教師なし学習処理システム１０３０の代わりに、スタック用教師あり／なし学習処理システム１０４０を備えた構成をとる。

【０２２０】

スタック用教師あり／なし学習処理システム１０４０は、スタック用教師あり／なし学習処理システム１０２０と同様の処理手段に解データベース２を追加した構成をとる。スタック用教師あり／なし学習処理システム１０４０は、言語解析処理について、文データベース５から生成した教師なしデータから解もしくは解候補と素性の集合との組を抽出し、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機械学習法により学習してその学習結果を記憶しておき、この学習結果を参照して第１素性追加部５２１または第２素性

追加部５２２から受け取った解もしくは解候補と素性の集合との組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（解析結果）と推定し、推定された解d_4を第１素性追加部５２１へまたは解d_4’を第２素性追加部５２２へ返却する手段である。

【０２２１】

スタック用教師あり／なし学習処理システム１０４０は、解d_4、解d_4’として、解と推定した解候補を出力するとともに、その解が正例もしくは負例であるかの情報や、正例もしくは負例である確率の情報などを出力することもできる。

【０２２２】

言語解析処理システム５８０の第１素性追加部５２１は、スタック用教師あり／なし学習処理システム１０４０から返却された解d_4を受け取り、”教師あり／なし学習の解析結果＝d_4”を素性として元の素性の集合に追加する。また、言語解析処理システム５８０の第２素性追加部５２２は、スタック用教師あり／なし学習処理システム１０４０から返却された解d_4’を受け取り、”教師あり／なし学習の解析結果＝d_4’”を素性として元の素性の集合に追加する。

【０２２３】

図２１に、第４の実施の形態における言語解析処理システムの別の構成例を示す。言語解析処理システム６００は、言語解析処理システム５００と同様の処理手段を備え、さらにスタック用教師あり／なし学習処理システム１０２０を備えた構成をとる。

【０２２４】

言語解析処理システム６００の第１素性追加部６１１は、解－素性対抽出部５０１から受け取った解と素性の集合との組から素性の集合のみをスタック用教師なし学習処理システム１０１０およびスタック用教師あり／なし学習処理システム１０２０へ渡し、スタック用教師なし学習処理システム１０１０から返却された解d_1およびスタック用教師あり／なし学習処理システム１０２０から返却された解d_2を受け取る。そして、”教師なし学習の解析結果＝d_1”および”教師あり／なし学習の解析結果＝d_2”を素性として元の素性の集合に追加する。

【０２２５】

また、言語解析処理システム６００の第２素性追加部６１２は、素性抽出部５０４から受け取った素性の集合をスタック用教師なし学習処理システム１０１０およびスタック用教師あり／なし学習処理システム１０２０へ渡し、スタック用教師なし学習処理システム１０１０から返却された解d_1’およびスタック用教師あり／なし学習処理システム１０２０から返却された解d_2’を受け取り、”教師なし学習の解析結果＝d_1’”および”教師あり／なし学習の解析結果＝d_2’”を素性として元の素性の集合に追加する。

【０２２６】

図２２に、第４の実施の形態における言語解析処理システムの別の構成例を示す。言語解析処理システム６５０は、言語解析処理システム５５０と同様の処理手段を備え、さらにスタック用教師あり／なし学習処理システム１０４０を備えた構成をとる。

【０２２７】

言語解析処理システム６５０の第１素性追加部６２１は、素性－解対・素性－解候補対抽出部５６１から受け取った解もしくは解候補と素性の集合との組をスタック用教師なし学習処理システム１０３０およびスタック用教師あり／なし学習処理システム１０４０へ渡し、スタック用教師なし学習処理システム１０３０から返却された解d_3およびスタック用教師あり／なし学習処理システム１０４０から返却された解d_4を受け取る。そして、”教師なし学習の解析結果＝d_3”および”教師あり／なし学習の解析結果＝d_4”を素性として元の素性の集合に追加する。

【０２２８】

また、言語解析処理システム６５０の第２素性追加部６２２は、素性－解候補抽出部５６４から受け取った解候補と素性の集合との組をスタック用教師なし学習処理システム１０３０およびスタック用教師あり／なし学習処理システム１０４０へ渡し、スタック用教師

なし学習処理システム１０３０から返却された解d_3’およびスタック用教師あり／なし学習処理システム１０４０から返却された解d_4’を受け取り、”教師なし学習の解析結果＝d_3’”および”教師あり／なし学習の解析結果＝d_4’”を素性として元の素性の集合に追加する。

【０２２９】

スタック用教師なし学習処理システム１０３０およびスタック用教師あり／なし学習処理システム１０４０は、解d_3、解d_3’、解d_4、解d_4’として、解と推定した解候補を出力するとともに、その解が正例もしくは負例であるかの情報や、正例もしくは負例である確率の情報などを出力することもできる。この場合には、受け取った解に含まれる情報の一部または全部が素性の集合に追加されるようにする。例えば、”教師なし学習の解析結果＝推定された解候補”、”教師なし学習の解析結果＝正例／負例”、または”教師なし学習の解析結果＝正例の確率／負例の確率”のような素性などの１つもしくは複数が元の素性の集合に追加される。

【０２３０】

すでに説明したように、教師なしデータは、教師ありデータと異なる性質を持つことから、単純に教師なしデータを教師ありデータに追加して機械学習を行うことが処理精度の改善に不十分である場合もある。本形態のようにスタッキング手法により教師なしデータによる機械学習と教師ありデータによる機械学習とを融合することで、これら双方の学習の利点を適切に利用することができ、解析処理の精度向上を図ることができたと思われる。

【０２３１】

最後に、従来技術による手法と本発明の手法の実施例を示す。実施例として受け身文・使役文から能動文への文変換処理における格変換処理を採用した。機械学習法としてサポートベクトルマシン法を採用した。また、京大コーパスを教師ありデータとして利用し、また、京大コーパスに含まれるの能動文のすべての格助詞（５３，１５７個）を教師なしデータとして利用した。図２３に、教師なしデータにおける変換後格助詞の分布を示す。

【０２３２】

さらに、実施例での処理精度の評価にも京大コーパスを用い、１０分割のクロスバリデーションにより評価を行った。

［参考文献６：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］

以下の方法を用いて格助詞の変換の実験を行なった。

【０２３３】

・教師あり学習の利用
・教師なし学習の利用
・教師あり／なし学習の利用
・スタッキング手法１：
教師なし学習の解析結果を素性に追加後、教師あり学習を行なう。

【０２３４】

・スタッキング手法２：
教師あり／なし学習の解析結果を素性に追加後、教師あり学習を行なう。

【０２３５】

・スタッキング手法３：
教師なし学習の解析結果と教師あり／なし学習の解析結果とを素性に追加後、教師あり学習を行なう。

【０２３６】

処理精度の評価結果を、以下に示す。処理精度は教師ありデータの事例数　４，６７１個のうち、どれだけ正解したかを意味する。

【０２３７】

・教師あり学習の利用＝８９．０６％
・教師なし学習の利用＝５１．１５％

・教師あり／なし学習の利用＝８７．０９％
・スタッキング手法１＝８９．４７％
・スタッキング手法２＝８９．５５％
・スタッキング手法３＝８９．５５％
教師あり学習方法を用いた処理の精度は、８９．０６％であった。これは、受け身文・使役文から能動文へ文変換における格助詞の変換処理を、機械学習法を用いて処理することにより、少なくともこの精度で実現できることを意味する。従来、機械学習法を用いた格助詞の変換処理はないので、本発明の実施例が示すこの精度は、本発明の格別な効果を示すものである。

【０２３８】
教師なし学習方法を用いた処理の精度は、５１．１５％と極めて低かった。解析対象である変換前格助詞の情報の欠如の影響が大きいと考えられる。

【０２３９】
また、教師あり／なし学習方法を用いた処理の精度も、教師あり学習方法を用いた処理の精度よりも低かった。教師なしデータは、教師ありデータとは異なる性質を持つため、教師なしデータの利用が精度低下を招いたと考えられる。

【０２４０】
すべてのスタッキング手法を用いた処理の精度は、教師あり学習方法を用いた処理の精度の精度を上回った。しかし、精度の向上は大きくない。そこで、二項検定を使って統計的検定を行なった結果、すべてのスタッキング手法が教師あり学習に対して有意水準０．０１で有意差を持った。このため、本発明における、教師なし学習の結果を素性に追加して利用する手法が、効果を持つことが確認できた。

【０２４１】
さらに、本発明の「教師あり学習を用いた処理」の精度との比較のため、従来技術の一つとして非特許文献４に記載された方法による処理を実施した。

【０２４２】
非特許文献４に記載された手法による格変換処理の精度はＦ値で３６％（再現率７５％、適合率２４％）であった。この従来技術による処理精度が低い理由は、与えられた文に辞書にない語が存在することである。そのような辞書に未定義の語を登録した後の処理の精度はＦ値で８３％（再現率９４％、適合率７４％）であった。なお、ここで精度をＦ値で示しているのは非特許文献４の手法での格変換は１つの入力に複数の変換結果を出力するためである。このように、すでに指摘したとおり既存の各フレーム辞書の不十分さの影響が大きいことがわかる。

【０２４３】
また、非特許文献４の手法による処理結果が文単位であるため、本発明による処理結果も文単位で集計した。このとき、本発明による処理では、文単位の精度は８５．５８％であった。ただし、ここでの文単位は用言が１つの文であり、複文など複数の文により構成されている文は用言が１つの文に分割してから精度の算出を行なった。

【０２４４】
本発明による処理の精度は、非特許文献４に示す手法で未知語などを辞書に登録した後の処理精度と同程度である。本発明では、解析対象となる情報について辞書への追加登録などは一切行なわずに８５％程度の精度を得ている。このことから、本発明による処理が、従来技術より高い精度で処理を行えることがわかる。

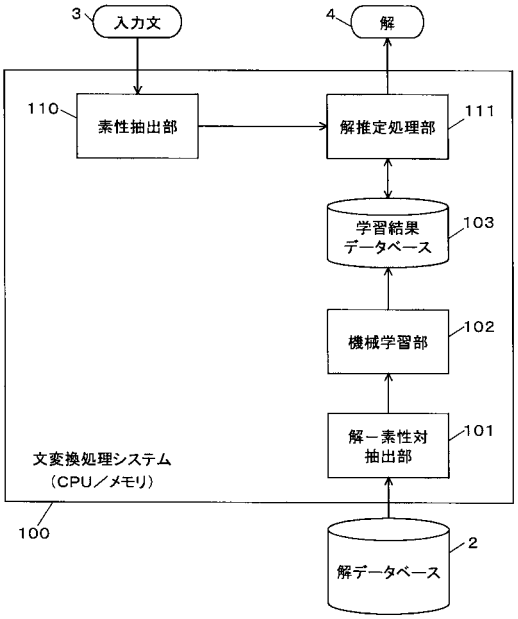
【０２４５】
以上、本発明をその実施の形態により説明したが、本発明はその主旨の範囲において種々の変形が可能であることは当然である。

【０２４６】
本発明の実施の形態では、主に受け身文、使役文から能動文への変換処理における格助詞の変換を扱った。しかし、本発明における機械学習部での分類先を能動文での格助詞から受け身文、使役文での格助詞とすることにより、能動文から受け身文、使役文への変換処

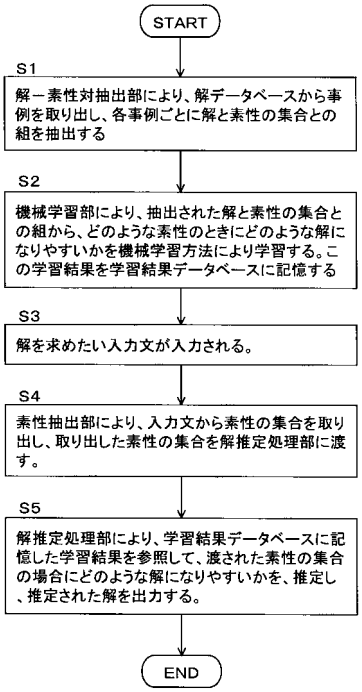
理についても本発明を適用することが可能である。
【０２４７】
また、本発明の実施の形態で言語解析処理として説明した解析処理以外にも、指示詞・代名詞・ゼロ代名詞などの照応解析、間接照応解析、「ＡのＢ」の意味解析、換喩解析などの種々の解析処理、文生成処理における格助詞生成処理、翻訳処理における格助詞生成処理などの処理についても本発明を適用することが可能である。
【０２４８】
また、本発明の各手段または機能または要素は、コンピュータにより読み取られ実行される処理プログラムとして実現することができる。また、本発明を実現する処理プログラムは、コンピュータが読み取り可能な、可搬媒体メモリ、半導体メモリ、ハ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ものである。
【０２４９】
【発明の効果】
以上説明したように、本発明により、教師なしデータを用いた機械学習の解析結果を素性に追加し、追加された素性を持つ教師ありデータを用いて機械学習を行なう新しい手法を実現した。これにより、教師なしデータと教師ありデータの双方の利点を用いた機械学習が実現でき、より高い精度の文変換処理を実現することが可能となった。
【０２５０】
特に本発明は、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、語句生成処理を含むようなきわめて広範囲の問題に適用することができる。これにより、実用性の高い言語解析処理システムを実現することができる。
【０２５１】
また、本発明により、日本語の受け身文・使役文から能動文へ変換処理における格助詞の変換を機械学習を用いて行う新しい手法を実現した。本発明により、従来に比べて高い精度で変換後格助詞の推定を行うことが可能となった。
【０２５２】
本発明を適用した受け身文・使役文から能動文への変換は、文生成処理、文言い換え処理、知識獲得システム、質問応答システムなどのコンピュータを用いた自然言語処理の数多くの分野で役に立つものである。
【図面の簡単な説明】
【図１】第１の実施の形態における文変換処理システムの構成例を示す図である。
【図２】第１の実施の形態における文変換処理システムの処理フローを示す図である。
【図３】タグ付きコーパスに記憶されている事例の例を示す図である。
【図４】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５】第１の実施の形態における文変換処理システムの別の構成例を示す図である。
【図６】第１の実施の形態において別の構成例をとる文変換処理システムの処理フローを示す図である。
【図７】教師なしデータを説明するための図である。
【図８】第２の実施の形態における文変換処理システムの構成例を示す図である。
【図９】教師なしデータ生成処理の処理フローを示す図である。
【図１０】第２の実施の形態における文変換処理システムの別の構成例を示す図である。
【図１１】第３の実施の形態における文変換処理システムの構成例を示す図である。
【図１２】第３の実施の形態における文変換処理システムの別の構成例を示す図である。
【図１３】第４の実施の形態における言語解析処理システムの構成例を示す図である。
【図１４】第４の実施の形態における言語解析処理システムの処理フローを示す図である。
【図１５】第４の実施の形態における言語解析処理システムの処理フローを示す図である。

【図１６】第４の実施の形態における言語解析処理システムの別の構成例を示す図である。
【図１７】第４の実施の形態における言語解析処理システムの別の構成例を示す図である。
【図１８】第４の実施の形態において別の構成例をとる言語解析処理システムの処理フローを示す図である。
【図１９】第４の実施の形態において別の構成例をとる言語解析処理システムの処理フローを示す図である。
【図２０】第４の実施の形態における言語解析処理システムの別の構成例を示す図である。
【図２１】第４の実施の形態における言語解析処理システムの別の構成例を示す図である。
【図２２】第４の実施の形態における言語解析処理システムの別の構成例を示す図である。
【図２３】実施例において教師なしデータにおける変換後格助詞の分布を示す図である。
【符号の説明】
１００，１５０，２００，２５０，３００，３５０　文変換処理システム
１０１，５０１　解－素性対抽出部
１０２，１６２，５０２，５６２　機械学習部
１０３，１６３，５０３，５６３　学習結果データベース
１１０，５０４　素性抽出部
１１１，１７１，５０５，５６５　解推定処理部
１６１，５６１　素性－解対・素性－解候補対抽出部
１７０，５６４　素性－解候補抽出部
２０１　問題表現相当部抽出部
２０２　問題表現情報記憶部
２０３　意味解析情報記憶部
２０４　問題構造変換部
２０５　教師なしデータ記憶部
５００，５４０，５５０，５８０，６００，６５０　言語解析処理システム
５１１，５２１，６１１，６２１　第１素性追加部
５１２，５２２，６１２，６２２　第２素性追加部
１０１０，１０３０　スタック用教師なし学習処理システム
１０２０，１０４０　スタック用教師あり／なし学習処理システム
２，６　解データベース
３　入力文
４　解
５　文データベース

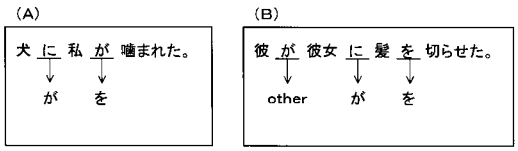
【図１】


3
入力文
4
解
110
素性抽出部
解推定処理部
111
学習結果
データベース
103
機械学習部
102
解－素性対
抽出部
101
文変換処理システム
（CPU／メモリ）
100
解データベース
2

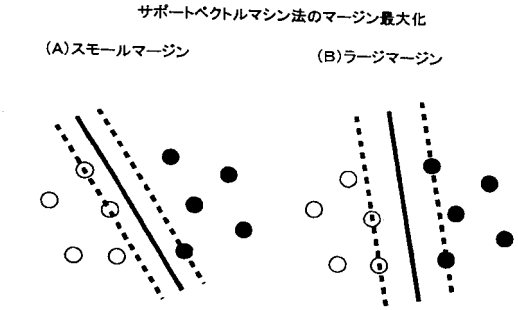
【図２】


START
S1
解－素性対抽出部により、解データベースから事例を取り出し、各事例ごとに解と素性の集合との組を抽出する
S2
機械学習部により、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方法により学習する。この学習結果を学習結果データベースに記憶する
S3
解を求めたい入力文が入力される。
S4
素性抽出部により、入力文から素性の集合を取り出し、取り出した素性の集合を解推定処理部に渡す。
S5
解推定処理部により、学習結果データベースに記憶した学習結果を参照して、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、推定された解を出力する。
END

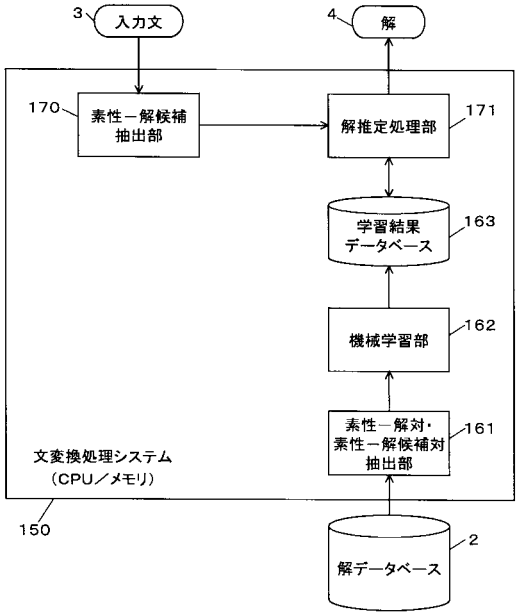
【図３】


(A)
犬 に 私 が 噛まれた。
が を
(B)
彼 が 彼女 に 髪 を 切らせた。
other が を

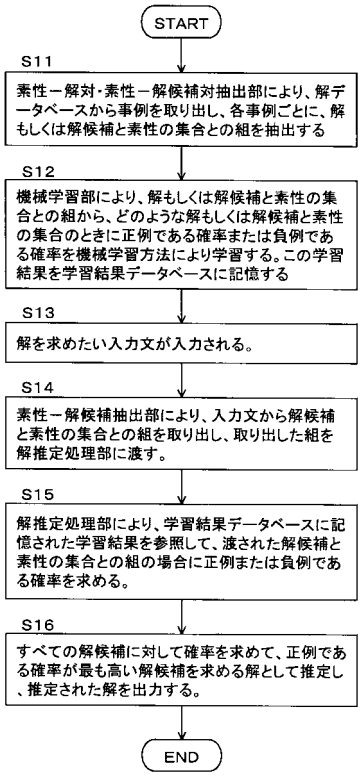
【図４】


サポートベクトルマシン法のマージン最大化
(A)スモールマージン
(B)ラージマージン

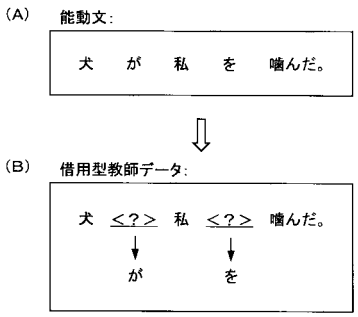
【図５】


3
入力文
4
解
170
素性－解候補
抽出部
171
解推定処理部
学習結果
データベース
163
機械学習部
162
素性－解対・
素性－解候補対
抽出部
161
文変換処理システム
（CPU／メモリ）
150
解データベース
2

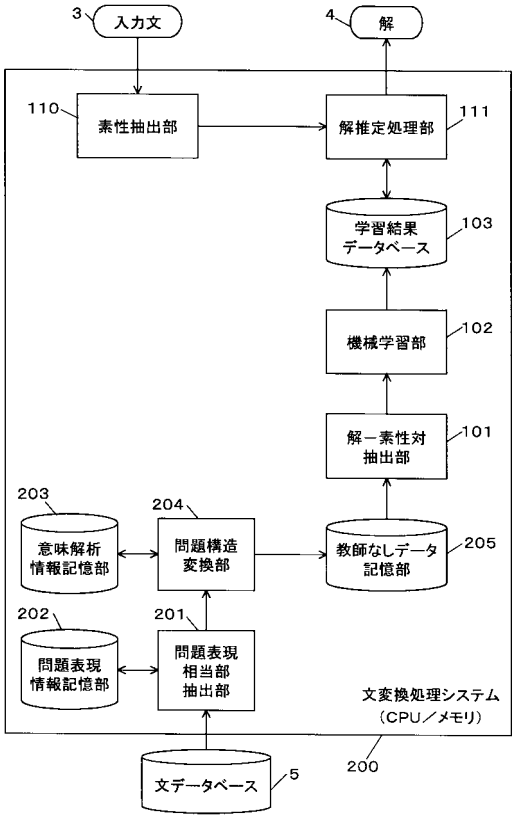
【図６】


START
S11
素性－解対・素性－解候補対抽出部により、解データベースから事例を取り出し、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
S12
機械学習部により、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機械学習方法により学習する。この学習結果を学習結果データベースに記憶する
S13
解を求めたい入力文が入力される。
S14
素性－解候補抽出部により、入力文から解候補と素性の集合との組を取り出し、取り出した組を解推定処理部に渡す。
S15
解推定処理部により、学習結果データベースに記憶された学習結果を参照して、渡された解候補と素性の集合との組の場合に正例または負例である確率を求める。
S16
すべての解候補に対して確率を求めて、正例である確率が最も高い解候補を求める解として推定し、推定された解を出力する。
END

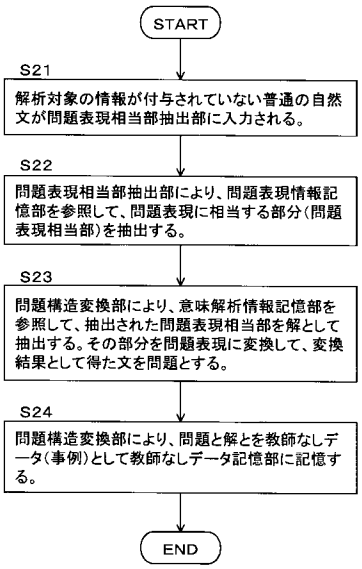
【図７】


(A)
能動文:
犬 が 私 を 噛んだ。
(B)
借用型教師データ:
犬 <?> 私 <?> 噛んだ。
が
を

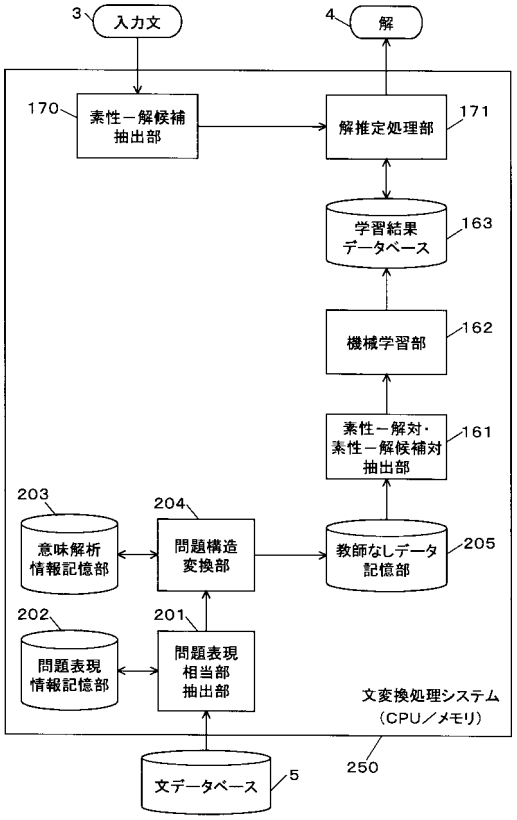
【図８】


3
入力文
4
解
110
素性抽出部
解推定処理部
111
学習結果
データベース
103
機械学習部
102
解－素性対
抽出部
101
203
意味解析
情報記憶部
204
問題構造
変換部
教師なしデータ
記憶部
205
202
問題表現
情報記憶部
201
問題表現
相当部
抽出部
文変換処理システム
（CPU／メモリ）
200
文データベース
5

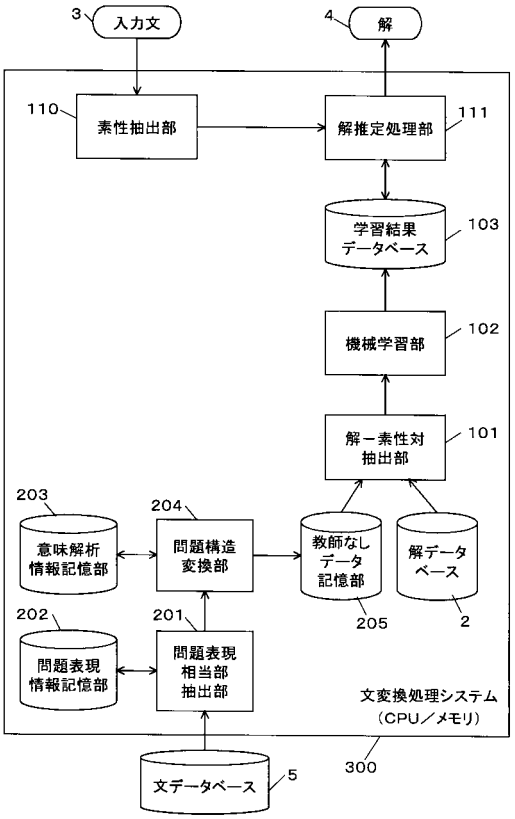
【図９】



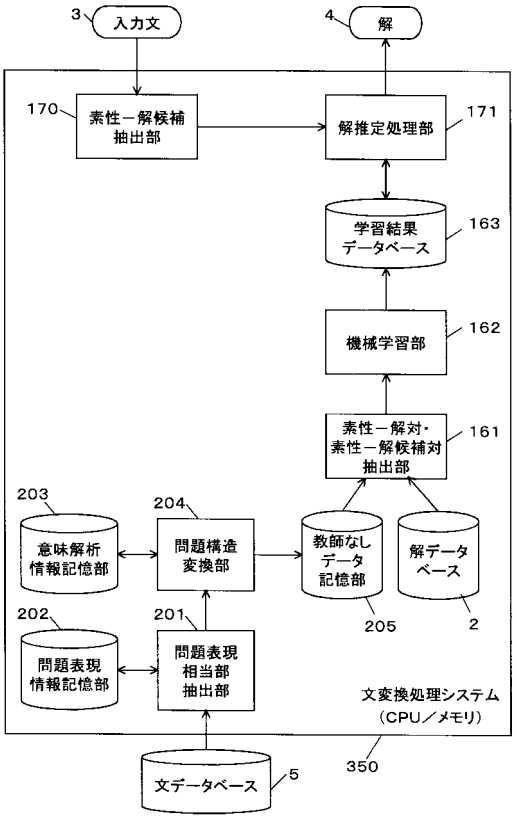
【図１０】



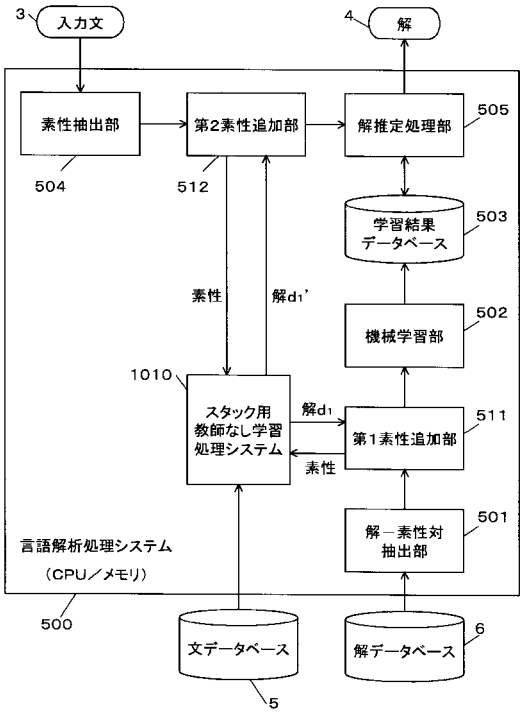
【図１１】



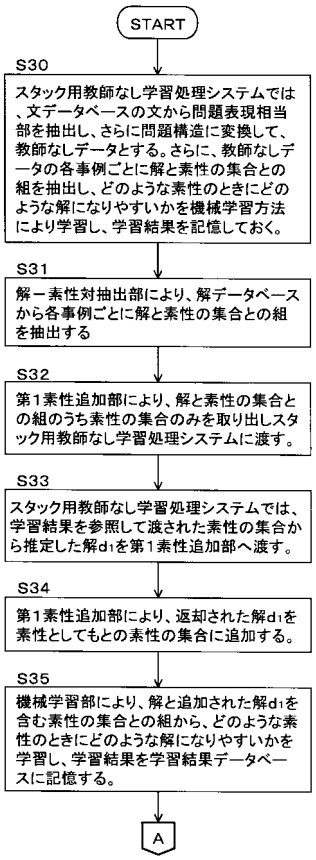
【図１２】



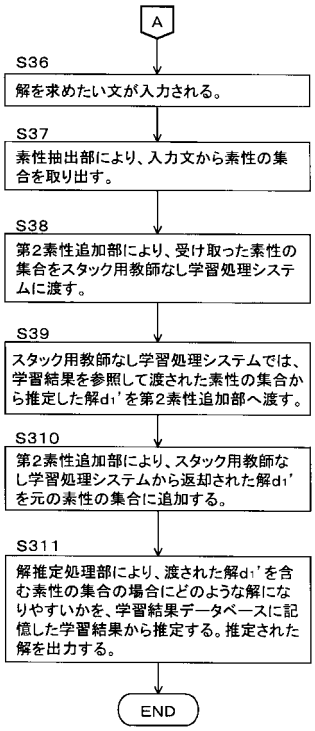
【図１３】



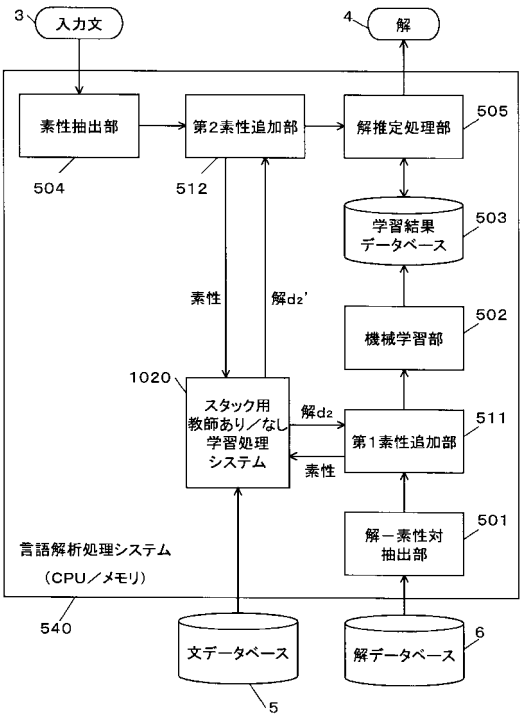
【図１４】



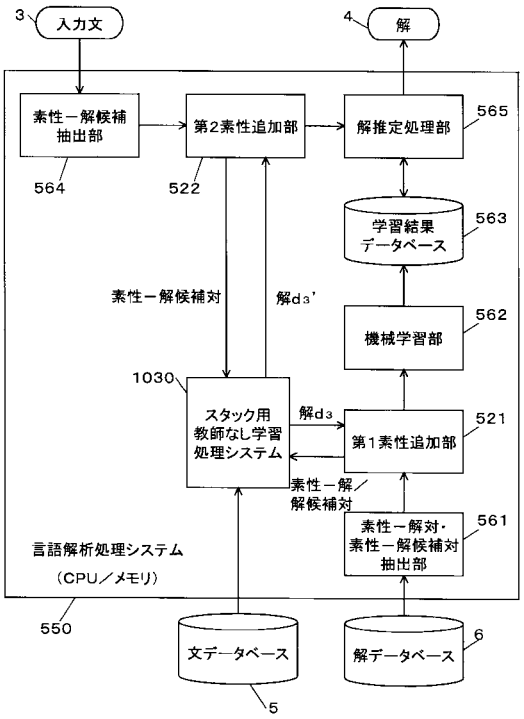
【図１５】



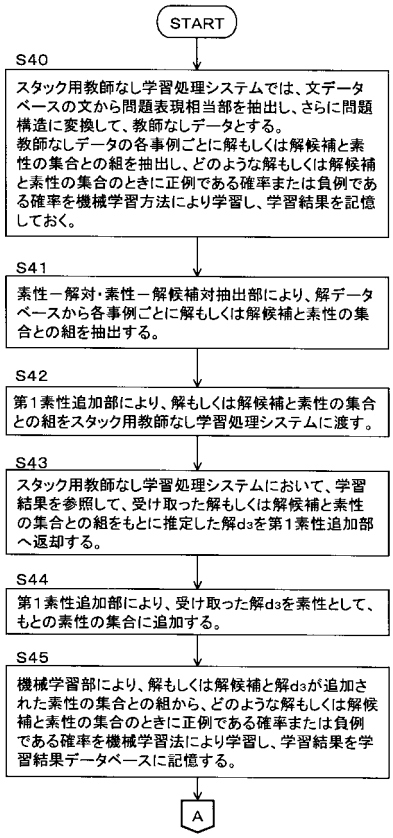
【図１６】



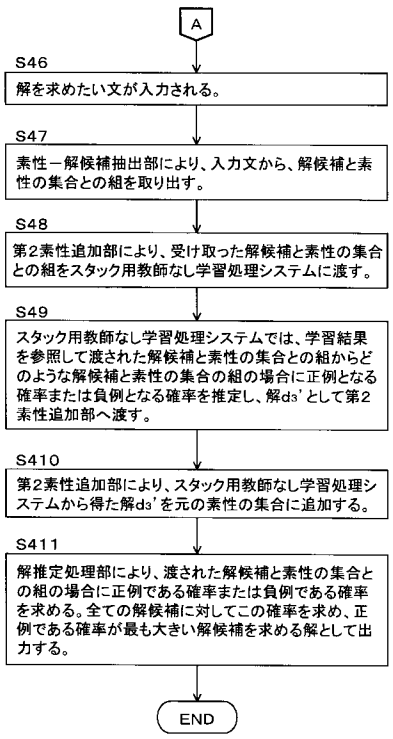
【図１７】



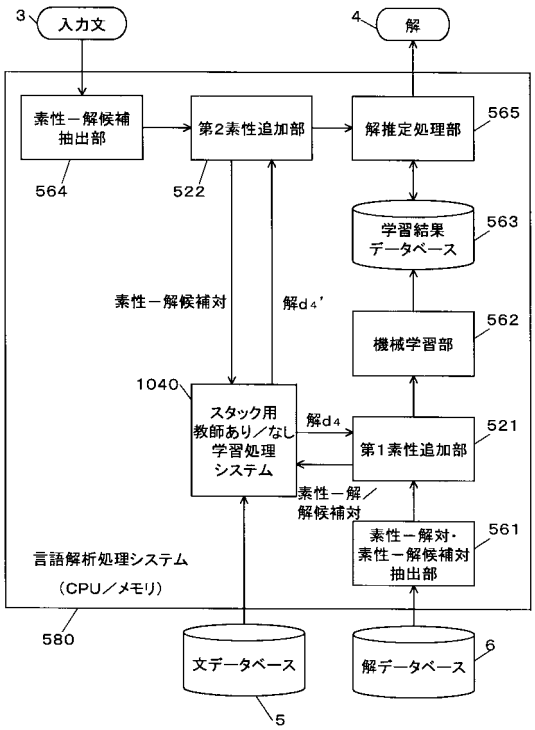
【図１８】



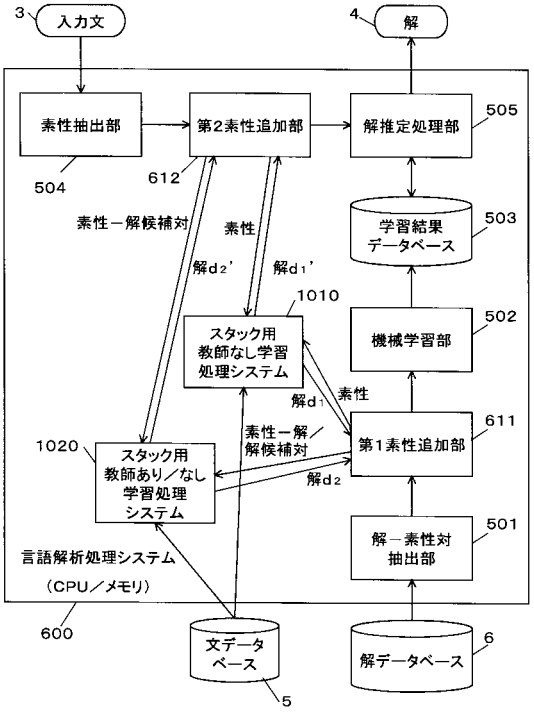
【図１９】



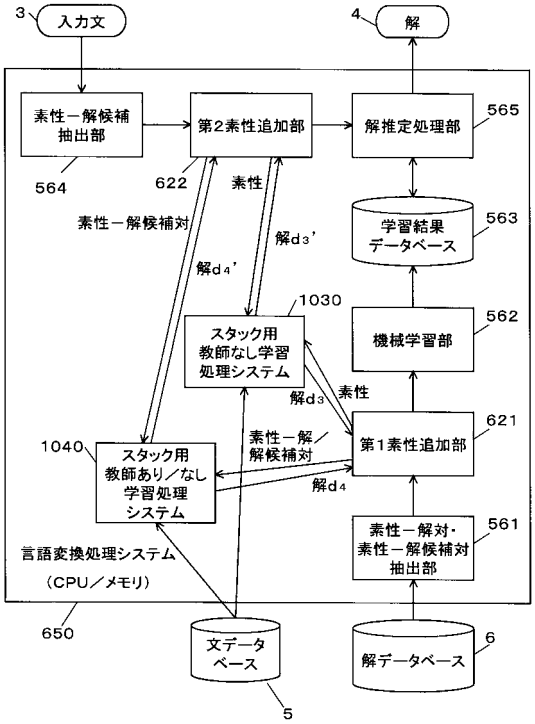
【図２０】



【図２１】



【図２２】



【図２３】

変換後格助詞	個数
を	14，535
に	13，148
が	9，792
と	7，849
で	5，654
から	1，490
まで	322
より	187
へ	177
にて	2
よ	1
合　計	53，157

フロントページの続き

(56)参考文献 村田真樹，機械学習手法を用いた日本語格解析，情報処理学会研究報告２００１－ＮＬ－１４４－１６，日本，社団法人情報処理学会，２００１年 7月17日，Ｖｏｌ．２００１，Ｎｏ．６９，ｐ．１１３－ｐ．１２０

(58)調査した分野(Int.Cl.，ＤＢ名)
G06F 17/21-17/28