

(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2004-171354
(P2004-171354A)

(43) 公開日 平成16年6月17日(2004.6.17)

(51) Int. Cl.⁷
G06F 17/27

F I
G06F 17/27

テーマコード(参考)
5B091

審査請求 有 請求項の数 10 O L (全 43 頁)

(21) 出願番号 (22) 出願日 特許法第30条第1項適用申請有り 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」において文書をもって発表	特願2002-337747 (P2002-337747) 平成14年11月21日 (2002.11.21)	(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 (74) 代理人 100119161 弁理士 重久 啓子 (74) 代理人 100097836 弁理士 福井 國敏 (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 Fターム(参考) 5B091 CA12 EA01
--	--	---

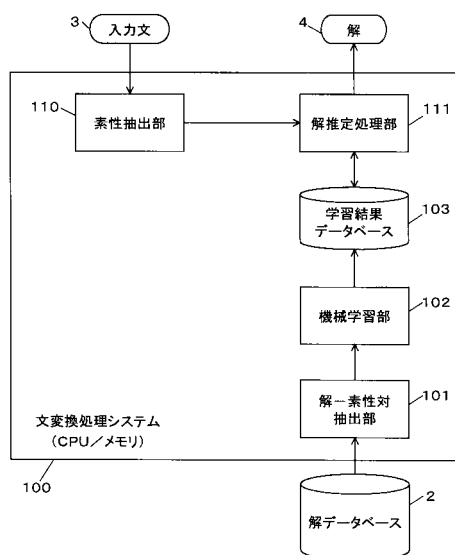
(54) 【発明の名称】 言語解析処理方法、文変換処理方法、言語解析処理システム、および文変換処理システム

(57) 【要約】

【課題】 電子化された文の変換処理を高精度に行える処理方法を提供する。

【解決手段】 解-素性対抽出部101は、解データベース2から事例データを取り出し、各事例データごとに解と素性の集合との組を抽出する。機械学習部102は、解と素性の集合との組を用いて、どのような素性の場合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、学習結果を学習結果データベース103に記憶する。素性抽出部110は、入力文3から素性の集合を抽出する。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から抽出された素性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。

【選択図】 図1



【特許請求の範囲】

【請求項 1】

スタック用処理システムと前記スタック用処理システムの処理結果を用いて処理を行うメイン処理システムとで構成される言語解析処理システムで機械学習法を用いて言語解析を処理する方法であって、

前記スタック用処理システムでは、

解析処理を行う場合に機械学習法が問題とする表現の情報を予め記憶した問題表現情報記憶部へアクセスし、

前記問題表現情報をもとに、当該問題についての情報が付与されていない事例データから前記問題表現情報に合致する問題表現相当部を抽出し、

10

前記問題表現相当部を解とし前記問題表現相当部を変換した前記事例データを問題とする構造を持つ教師なしデータに変換し、

前記教師なしデータごとに解と素性の集合との組を抽出し、

前記解と素性の集合との組から、どのような素性の場合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、

前記学習の結果を学習結果データベースに記憶しておき、

前記メイン処理システムから事例データから抽出された第 1 の素性の集合または処理対象として入力された事例データから抽出された第 2 の素性の集合を取得し、

前記学習結果データベースを参照して、前記第 1 の素性の集合または前記第 2 の素性の集合から、どのような素性の場合にどのような解になりやすいかを推定し、

20

前記第 1 の素性の集合に対する第 1 の処理結果または第 2 の素性の集合に対する第 2 の処理結果を前記メイン処理システムへ送出し、

前記メイン処理システムでは、

解析処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースにアクセスし、

前記解データベースから前記事例データを取り出し、前記事例データごとに解と素性の集合との組を抽出し、

前記解と素性の集合との組から素性の集合のみを取り出した第 1 の素性の集合を前記スタック用処理システムへ送出し、

前記スタック用処理システムから前記第 1 の処理結果を受け取り、

30

前記第 1 の処理結果を素性として前記事例データから抽出した第 1 の素性の集合に追加し、

前記解と前記処理結果を追加した第 1 の素性の集合との組から、どのような素性の場合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、

前記学習の結果を学習結果データベースに記憶し、

処理対象として入力された事例データから素性の集合を抽出し、

前記入力事例データの第 2 の素性の集合を前記スタック用処理システムへ送出し、

前記スタック用処理システムから出力された第 2 の処理結果を受け取り、

前記第 2 の処理結果を素性として前記入力事例データの第 2 の素性の集合に追加し、

前記学習結果データベースを参照して、前記処理結果を追加した前記第 2 の素性の集合から、どのような素性の場合にどのような解になりやすいかを推定する

40

ことを特徴とする言語解析処理方法。

【請求項 2】

スタック用処理システムと前記スタック用処理システムの処理結果を用いて処理を行うメイン処理システムとで構成される言語解析処理システムで機械学習法を用いて言語解析を処理する方法であって、

前記スタック用処理システムでは、

解析処理を行う場合に機械学習法が問題とする表現の情報を予め記憶した問題表現情報記憶部へアクセスし、

前記問題表現情報をもとに、当該問題についての情報が付与されていない事例データから

50

前記問題表現情報に合致する問題表現相当部を抽出し、
前記問題表現相当部を解とし前記問題表現相当部を変換した前記事例データを問題とする
構造を持つ教師なしデータに変換し、
前記教師なしデータごとに解もしくは解候補と素性の集合との組を抽出し、
前記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
合の場合に正例であるかの確率または負例であるかの確率を機械学習法により学習し、
前記学習の結果を学習結果データベースに記憶しておき、
前記メイン処理システムから、事例データから抽出された解もしくは解候補と素性の集合
の第1の組または処理対象として入力された事例データから抽出された解候補と素性の集
合の第2の組を取得し、
前記学習結果データベースを参照して、前記第1の組または前記第2の組から、どのよう
な解候補と素性の集合の場合に正例であるかの確率または負例であるかの確率を求め、す
べての解候補の中から正例である確率が最も高い解候補を解と推定する処理を行い、
前記第1の組に対する第1の処理結果または第2の組に対する第2の処理結果を前記メイ
ン処理システムへ送出し、
前記メイン処理システムでは、
解析処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶
する解データベースにアクセスし、
前記解データベースから前記事例データを取り出し、前記事例データごとに解もしくは解
候補と素性の集合との第1の組を抽出し、
前記第1の組を前記スタック用処理システムへ送出し、
前記スタック用処理システムから前記第1の処理結果を受け取り、
前記第1の処理結果を前記第1の組の素性の集合に追加し、
前記第1の処理結果を追加した前記第1の組から、どのような解もしくは解候補と素性の
集合の場合に正例であるかの確率または負例であるかの確率を機械学習法により学習し、
前記学習の結果を学習結果データベースに記憶し、
処理対象として入力された事例データから解候補と素性の集合の第2の組を抽出し、
前記第2の組を前記スタック用処理システムへ送出し、
前記スタック用処理システムから第2の処理結果を受け取り、
前記第2の処理結果を素性として前記第2の組の素性の集合に追加し、
前記学習結果データベースを参照して、前記第2の処理結果を追加した前記第2の組から
、どのような解候補と素性の集合の場合に正例であるかの確率または負例であるかの確率
を求め、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る解として推
定する
ことを特徴とする言語解析処理方法。

10

20

30

40

50

【請求項3】

請求項1または請求項2のいずれかに記載の言語解析処理方法において、
前記スタック用処理システムおよび前記メイン処理システムでは、受け身文もしくは使役
文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
ことを特徴とする言語解析処理方法。

【請求項4】

コンピュータを用いて電子化された受け身文または使役文を能動文へ変換処理する文変換
処理方法であって、
受け身文または使役文から能動文への変換処理を行う場合に機械学習法が扱う問題につ
いての解が付与された事例データを記憶する解データベースへアクセスし、
前記解データベースから前記事例データを取り出し、前記事例データごとに解と素性の集
合との組を抽出し、
前記解と素性の集合との組から、どのような素性の場合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、
前記学習の結果を学習結果データベースに記憶し、

解析対象として入力された電子化された文から素性の集合を抽出し、
前記学習結果データベースを参照して、前記入力文から抽出された素性の場合にどのような解になりやすいかを推定することを特徴とする文変換処理方法。

【請求項 5】

コンピュータを用いて電子化された受け身文または使役文を能動文へ変換処理する文変換処理方法であって、

受け身文または使役文から能動文への変換処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースへアクセスし、

前記解データベースから前記事例データを取り出し、前記事例データごとに解もしくは解候補と素性の集合との組を抽出し、

前記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場合に正例であるかの確率または負例であるかの確率を機械学習法により学習し、

前記学習の結果を学習結果データベースに記憶し、

解析対象として入力された電子化された文から解候補と素性の集合との組を抽出し、

前記学習結果データベースを参照して、前記入力文から抽出され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の場合に正例であるかの確率または負例であるかの確率を求め、すべての解候補の中から正例である確率が最も高い解候補を解と推定する

ことを特徴とする文変換処理方法。

【請求項 6】

スタック用処理システムと前記スタック用処理システムの処理結果を用いて処理を行うメイン処理システムとで構成される機械学習法を用いて言語解析を処理するシステムであって、

前記スタック用処理システムは、

解析処理を行う場合に機械学習法が問題とする表現の情報を予め記憶した問題表現情報記憶手段と、

前記問題表現情報をもとに、当該問題についての情報が付与されていない事例データから前記問題表現情報に合致する問題表現相当部を抽出する処理手段と、前記問題表現相当部を解とし前記問題表現相当部を変換した前記事例データを問題とする構造を持つ教師なし

データに変換する処理手段と、

前記教師なしデータごとに解と素性の集合との組を抽出する処理手段と、

前記解と素性の集合との組から、どのような素性の場合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する処理手段と、

前記学習の結果を記憶する学習結果データベースと、

前記メイン処理システムから事例データから抽出された第 1 の素性の集合または処理対象として入力された事例データから抽出された第 2 の素性の集合を取得する処理手段と、

前記学習結果データベースを参照して、前記第 1 の素性の集合または前記第 2 の素性の集合から、どのような素性の場合にどのような解になりやすいかを推定する処理手段と、

前記第 1 の素性の集合に対する第 1 の処理結果または第 2 の素性の集合に対する第 2 の処理結果を前記メイン処理システムへ送出手段とを備え、

前記メイン処理システムは、

解析処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースと、

前記解データベースから前記事例データを取り出し、前記事例データごとに解と素性の集合との組を抽出する処理手段と、

前記解と素性の集合との組から素性の集合のみを取り出した第 1 の素性の集合を前記スタック用処理システムへ送出手段と、

前記スタック用処理システムから前記第 1 の処理結果を受け取る処理手段と、前記第 1 の処理結果を素性として前記事例データから抽出した第 1 の素性の集合に追加する処理手段

と、

前記解と前記処理結果を追加した第1の素性の集合との組から、どのような素性の場合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する処理手段と、

前記学習の結果を記憶する学習結果データベースと、

処理対象として入力された事例データから素性の集合を抽出する処理手段と、

前記入力事例データの第2の素性の集合を前記スタック用処理システムへ送出手段と、

前記スタック用処理システムから出力された第2の処理結果を受け取る処理手段と、

前記第2の処理結果を素性として前記入力事例データの第2の素性の集合に追加する処理手段と、

前記学習結果データベースを参照して、前記処理結果を追加した前記第2の素性の集合から、どのような素性の場合にどのような解になりやすいかを推定する処理手段とを備えることを特徴とする言語解析処理システム。

【請求項7】

スタック用処理システムと前記スタック用処理システムの処理結果を用いて処理を行うメイン処理システムとで構成され、機械学習法を用いて言語解析を処理するシステムであって、

前記スタック用処理システムは、

解析処理を行う場合に機械学習法が問題とする表現の情報を予め記憶した問題表現情報記憶手段と、

前記問題表現情報をもとに、当該問題についての情報が付与されていない事例データから前記問題表現情報に合致する問題表現相当部を抽出する処理手段と、

前記問題表現相当部を解とし前記問題表現相当部を変換した前記事例データを問題とする構造を持つ教師なしデータに変換する処理手段と、

前記教師なしデータ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処理手段と、

前記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場合に正例であるかの確率または負例であるかの確率を機械学習法により学習する処理手段と、

前記学習の結果を記憶する学習結果データベースと、

前記メイン処理システムから、事例データから抽出された解もしくは解候補と素性の集合の第1の組または処理対象として入力された事例データから抽出された解候補と素性の集合の第2の組を取得する処理手段と、

前記学習結果データベースを参照して、前記第1の組または前記第2の組から、どのような解候補と素性の集合の場合に正例であるかの確率または負例であるかの確率を求め、すべての解候補の中から正例である確率が最も高い解候補を解と推定する処理手段と、

前記第1の組に対する第1の処理結果または第2の組に対する第2の処理結果を前記メイン処理システムへ送出手段とを備え、

前記メイン処理システムは、

解析処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースと、

前記解データベースから前記事例データを取り出し、前記事例データごとに解もしくは解候補と素性の集合との第1の組を抽出する処理手段と、

前記第1の組を前記スタック用処理システムへ送出手段と、

前記スタック用処理システムから前記第1の処理結果を受け取る処理手段と、

前記第1の処理結果を前記第1の組の素性の集合に追加する処理手段と、

前記第1の処理結果を追加した前記第1の組から、どのような解もしくは解候補と素性の集合の場合に正例であるかの確率または負例であるかの確率を機械学習法により学習する処理手段と、

前記学習の結果を記憶する学習結果データベースと、

処理対象として入力された事例データから解候補と素性の集合の第2の組を抽出する処理

10

20

30

40

50

手段と、

前記第 2 の組を前記スタック用処理システムへ送出する処理手段と、

前記スタック用処理システムから第 2 の処理結果を受け取る処理手段と、

前記第 2 の処理結果を素性として前記第 2 の組の素性の集合に追加する処理手段と、

前記学習結果データベースを参照して、前記第 2 の処理結果を追加した前記第 2 の組から、どのような解候補と素性の集合の場合に正例であるかの確率または負例であるかの確率を求め、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る解として推定する処理手段とを備える

ことを特徴とする言語解析処理システム。

【請求項 8】

10

請求項 6 または請求項 7 のいずれかに記載の言語解析処理システムにおいて、

前記スタック用処理システムおよび前記メイン処理システムでは、受け身文もしくは使役文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

ことを特徴とする言語解析処理システム。

【請求項 9】

コンピュータを用いて電子化された受け身文または使役文を能動文へ変換処理する文変換処理システムであって、

受け身文または使役文から能動文への変換処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースと、

前記解データベースから前記事例データを取り出し、前記事例データごとに解と素性の集合との組を抽出する処理手段と、

前記解と素性の集合との組から、どのような素性の場合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する処理手段と、

前記学習の結果を記憶する学習結果データベースと、

解析対象として入力された電子化された文から素性の集合を抽出する処理手段と、

前記学習結果データベースを参照して、前記入力文から抽出された素性の場合にどのような解になりやすいかを推定する処理手段とを備える

ことを特徴とする文変換処理システム。

【請求項 10】

コンピュータを用いて電子化された受け身文または使役文を能動文へ変換処理する文変換処理システムであって、

受け身文または使役文から能動文への変換処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースと、

前記解データベースから前記事例データを取り出し、前記事例データ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処理手段と、

前記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場合に正例であるかの確率または負例であるかの確率を機械学習法により学習する処理手段と、

前記学習の結果を記憶する学習結果データベースと、

解析対象として入力された電子化された文から解候補と素性の集合との組を抽出する処理手段と、

前記学習結果データベースを参照して、前記入力文から抽出され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の場合に正例であるかの確率または負例であるかの確率を求め、すべての解候補の中から正例である確率が最も高い解候補を解と推定する処理手段とを備える

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

50

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、

10

【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行う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」とであると解析する。

20

【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。

30

(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

40

【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。

50

【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。

10

【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。

20

【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

30

【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

40

【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った効果を持つ。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

50

。

【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。

10

【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つたり、誤変換する確率が高かつたりするという問題が生じていた。

20

【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。

30

【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、スタック用処理システムと前記スタック用処理システムの処理結果を用いて処理を行うメイン処理システムとで構成される言語解析処理システムで機械学習法を用いて言語解析を処理する方法であって、前記スタック用処理システムでは、解析処理を行う場合に機械学習法が問題とする表現の情報を予め記憶した問題表現情報記憶部へアクセスし、前記問題表現情報をもとに、当該問題についての情報が付与されていない事例データから前記問題表現情報に合致する問題表現相当部を抽出し、前記問題表現相当部を解とし前記問題表現相当部を変換した前記事例データを問題とする構造を持つ教師なしデータに変換する。そして、前記教師なしデータごとに解と素性の集合との組を抽出し、前記解と素性の集合との組から、どのような素性の場合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、前記学習の結果を学習結果データベースに記憶しておく。その後、前記メイン処理システムから事例データから抽出された第1の素性の集合または処理対象として入力された事例データから抽出された第2の素性の集合を取得し、前記学習結果データベ

40

50

ースを参照して、前記第1の素性の集合または前記第2の素性の集合から、どのような素性の場合にどのような解になりやすいかを推定し、前記第1の素性の集合に対する第1の処理結果または第2の素性の集合に対する第2の処理結果を前記メイン処理システムへ送

【0025】

前記メイン処理システムでは、解析処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースにアクセスし、前記解データベースから前記事例データを取り出し、前記事例データごとに解と素性の集合との組を抽出し、前記解と素性の集合との組から素性の集合のみを取り出した第1の素性の集合を前記スタック用処理システムへ送

10

【0026】

また、本発明は、スタック用処理システムと前記スタック用処理システムの処理結果を用いて処理を行うメイン処理システムとで構成される言語解析処理システムで機械学習法を用いて言語解析を処理する方法であって、前記スタック用処理システムでは、解析処理を行う場合に機械学習法が問題とする表現の情報を予め記憶した問題表現情報記憶部へアクセスし、前記問題表現情報をもとに、当該問題についての情報が付与されていない事例データから前記問題表現情報に合致する問題表現相当部を抽出し、前記問題表現相当部を解とし前記問題表現相当部を変換した前記事例データを問題とする構造を持つ教師なしデータに変換し、前記教師なしデータごとに解もしくは解候補と素性の集合との組を抽出し、前記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場合に正例であるかの確率または負例であるかの確率を機械学習法により学習し、前記学習の結果を学習結果データベースに記憶しておく。その後、前記メイン処理システムから、事例データから抽出された解もしくは解候補と素性の集合の第1の組または処理対象として入力された事例データから抽出された解候補と素性の集合の第2の組を取得し、前記学習結果データベースを参照して、前記第1の組または前記第2の組から、どのような解候補と素性の集合の場合に正例であるかの確率または負例であるかの確率を求め、すべての解候補の中から正例である確率が最も高い解候補を解と推定する処理を行い、前記第1の組に対する第1の処理結果または第2の組に対する第2の処理結果を前記メイン処理システムへ送

20

30

【0027】

前記メイン処理システムでは、解析処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースにアクセスし、前記解データベースから前記事例データを取り出し、前記事例データごとに解もしくは解候補と素性の集合との第1の組を抽出し、前記第1の組を前記スタック用処理システムへ送

40

50

集合に追加し、前記学習結果データベースを参照して、前記第2の処理結果を追加した前記第2の組から、どのような解候補と素性の集合の場合に正例であるかの確率または負例であるかの確率を求め、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る解として推定する。

【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。

10

【0029】

また、本発明は、コンピュータを用いて電子化された受け身文または使役文を能動文へ変換処理する文変換処理方法であって、受け身文または使役文から能動文への文変換処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースへアクセスし、前記解データベースから前記事例データを取り出し、前記事例データごとに解と素性の集合との組を抽出し、前記解と素性の集合との組から、どのような素性の場合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、前記学習の結果を学習結果データベースに記憶する。そして、解析対象として入力された電子化された文から素性の集合を抽出し、前記学習結果データベースを参照して、前記入力文から抽出された素性の場合にどのような解になりやすいかを推定する。

20

【0030】

また、本発明は、コンピュータを用いて電子化された受け身文または使役文を能動文へ変換処理する文変換処理方法であって、受け身文または使役文から能動文への変換処理を行う場合に機械学習法が扱う問題についての解が付与された事例データを記憶する解データベースへアクセスし、前記解データベースから前記事例データを取り出し、前記事例データごとに解もしくは解候補と素性の集合との組を抽出し、前記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場合に正例であるかの確率または負例であるかの確率を機械学習法により学習し、前記学習の結果を学習結果データベースに記憶する。そして、解析対象として入力された電子化された文から解候補と素性の集合との組を抽出し、前記学習結果データベースを参照して、前記入力文から抽出され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の場合に正例であるかの確率または負例であるかの確率を求め、すべての解候補の中から正例である確率が最も高い解候補を解と推定する。

30

【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。

40

【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

50

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。

10

20

30

40

50

【 0 0 4 5 】

図 2 に、文変換処理システム 1 0 0 の処理フローを示す。

【 0 0 4 6 】

ステップ S 1 : 解 - 素性対抽出部 1 0 1 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。

【 0 0 4 7 】

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。

10

【 0 0 4 8 】

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。

【 0 0 4 9 】

20

- 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。

30

【 0 0 5 0 】

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 1 0 1 は、抽出した素性の集合を機械学習部 1 0 2 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。

【 0 0 5 1 】

ステップ S 2 : 機械学習部 1 0 2 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 1 0 3 に記憶する。

40

【 0 0 5 2 】

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。

50

【0053】

また、事例「へびに噛まれた。 が」から抽出された、

- ・推定すべき格にある体言 n の単語 = へび、
- ・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 10
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、解析すべき入力を与えられたときに、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 ($1 \leq j \leq k$) 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 20
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ある。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 30
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, Tiny SVM: Support Vector Machines, (http://cl.aist-nara.ac.jp/taku-ku/software/Tiny_SVM/index.html, 2000 40
)]

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、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 50

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】

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

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

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

10

20

【0061】

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

$$\operatorname{sgn}(x) = \begin{cases} 1 & (x \geq 0) \\ -1 & (\text{otherwise}) \end{cases} \quad (2)$$

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。

【0062】

【数2】

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

$$0 \leq \alpha_i \leq C \quad (i = 1, \dots, l) \quad (4)$$

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

30

40

【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】

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \quad (6)$$

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

50

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、 $x_i > 0$ 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る用になる。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、Tiny SVM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」、「へびに噛まれた。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」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。

10

20

30

40

50

【 0 0 8 5 】

〔 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 。

【 0 0 8 6 】

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 。

10

【 0 0 8 7 】

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 。

20

【 0 0 8 8 】

「 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 。

【 0 0 8 9 】

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 。

30

【 0 0 9 0 】

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 2 0 0 は 、 C P U お よ び メ モ リ か ら な り 、 問 題 表 現 相 当 部 抽 出 部 2 0 1 、 問 題 表 現 情 報 記 憶 部 2 0 2 、 意 味 解 析 情 報 記 憶 部 2 0 3 、 問 題 構 造 変 換 部 2 0 4 、 教 師 な し デ ー タ 記 憶 部 2 0 5 、 解 - 素 性 対 抽 出 部 1 0 1 、 機 械 学 習 部 1 0 2 、 学 習 結 果 デ ー タ ベ ー ス 1 0 3 、 素 性 抽 出 部 1 1 0 、 解 推 定 処 理 部 1 1 1 、 お よ び 文 デ ー タ ベ ー ス 5 を 備 え る 。

【 0 0 9 1 】

問 題 表 現 相 当 部 抽 出 部 2 0 1 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 2 0 2 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 。

40

【 0 0 9 2 】

こ こ で は 、 問 題 表 現 情 報 記 憶 部 2 0 2 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 。

【 0 0 9 3 】

問 題 構 造 変 換 部 2 0 4 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 2 0 3 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 2 0 5 に 記 憶 す

50

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、

10

20

30

40

50

- ・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。 10

【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。 20

【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。 30

【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。 40

【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。

10

【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。

20

【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。

30

【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。

40

[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

50

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「d₁」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「d₁」を追加し、リスト{a, b, c, "教師なし学習の解析結果 = d₁"}を新しい素性の集合として機械学習を行なう。 10

【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「d₂」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「d₂」を追加し、リスト{a, b, c, "教師あり/なし学習の解析結果 = d₂"}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「d₁」および「d₂」を追加し、リスト{a, b, c, "教師なし学習の解析結果 = d₁", "教師あり/なし学習の解析結果 = d₂"}を新しい素性の集合として機械学習を行なう。 20

【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だけの場合には、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。 30

【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。 40

【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 50

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。

10

【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。

20

【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。

30

【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。

40

【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、 $\{a, b, c, \text{“教師なし学習の解析結果} = d_1\text{”}\}$ 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】

50

ステップ S 3 6 : 解を求めたい文が素性抽出部 5 0 4 に入力される。

【 0 1 4 0 】

ステップ S 3 7 : 素性抽出部 5 0 4 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 5 1 2 へ渡す。

【 0 1 4 1 】

ステップ S 3 8 : 第 2 素性追加部 5 1 2 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 1 0 1 0 へ渡される。

【 0 1 4 2 】

ステップ S 3 9 : スタック用教師なし学習処理システム 1 0 1 0 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 5 1 2 へ渡す。 10

【 0 1 4 3 】

ステップ S 3 1 0 : 第 2 素性追加部 5 1 2 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、機械学習部 5 0 2 に渡される素性の集合は、 $\{ a, b, c, \text{” 教師なし学習の解析結果} = d_1 \text{”} \}$ となり、この素性の集合が解推定処理部 5 0 5 へ渡される。

【 0 1 4 4 】

ステップ S 3 1 1 : 解推定処理部 5 0 5 により、学習結果データベース 5 0 3 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。 20

【 0 1 4 5 】

以下に、具体的な処理を例として言語解析処理システム 5 0 0 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 5 0 0 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。

【 0 1 4 6 】

言語解析処理システム 5 0 0 のスタック用教師なし学習処理システム 1 0 1 0 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、 30

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。

【 0 1 4 7 】

- ・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。

【 0 1 4 8 】

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、 40

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。

【 0 1 4 9 】

- ・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。

【 0 1 5 0 】

その後、解 - 素性対抽出部 5 0 1 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。

【0 1 5 1】

- ・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 5 1 1 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 1 0 1 0 へ渡す。スタック用教師なし学習処理システム 1 0 1 0 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 5 1 1 へ返却する。

10

【0 1 5 2】

次に、第 1 素性追加部 5 1 1 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。

【0 1 5 3】

- ・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 5 0 2 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 5 0 3 に記憶する。

20

【0 1 5 4】

その後、解を求めたい文が素性抽出部 5 0 4 に入力される。素性抽出部 5 0 4 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「へびに噛まれる。」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 5 1 2 へ渡す。

【0 1 5 5】

- ・推定すべき格にある体言 n = へび、
- ・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 5 1 2 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 1 0 1 0 へ渡される。スタック用教師なし学習処理システム 1 0 1 0 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 5 1 2 へ返却する。

30

【0 1 5 6】

第 2 素性追加部 5 1 2 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。

【0 1 5 7】

- ・推定すべき格にある体言 n = へび、
- ・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 5 0 5 へ渡される。解推定処理部 5 0 5 により、学習結果データベース 5 0 3 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。

40

【0 1 5 8】

ここでは、スタック用教師なし学習処理システム 1 0 1 0 から返却された解析結果「が」

50

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。

10

【0160】

さらに、解推定処理部505は、素性の集合の情報が多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。

20

【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。

30

【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。

40

【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。

50

【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)

10

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。

20

【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。

30

【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、

40

- ・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ます。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

50

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 $n = \text{みかん}$ 、
- ・生成すべき格が修飾する用言 $v = \text{食べる}$ 、
- ・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。

10

【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 $n = \text{みかん}$ 、
- ・生成すべき格が修飾する用言 $v = \text{食べる}$ 、
- ・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ため、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。

20

【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。

30

【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」

40

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

50

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」、

10

…、

- ・「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。

20

【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」、

30

…、

- ・「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」、

40

…、

- ・「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。

50

【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 ')

10

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。

20

【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。

30

【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。

40

【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1

50

7 1 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。 10

【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。 20

【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。 30

【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。 40

【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。 50

【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つもしくは複数がある元の素性の集合に追加される。

10

【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。

20

【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。

30

【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。

40

【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】

50

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。 10

【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。 20

【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。 30

【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。 40

【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 50

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。

10

【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

20

【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。

30

【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、pp 115 - 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。

40

【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 :

50

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%

10

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。

20

【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。

30

【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。

40

【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、

50

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、文言文換え処理、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。

10

20

30

40

50

- 【図 1 0】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 1 1】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 1 2】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 1 3】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 1 4】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 1 5】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 1 6】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 1 7】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 1 8】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 1 9】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 2 0】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 2 1】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 2 2】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 2 3】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】
- 1 0 0 , 1 5 0 , 2 0 0 , 2 5 0 , 3 0 0 , 3 5 0 文変換処理システム
- 1 0 1 , 5 0 1 解 - 素性対抽出部
- 1 0 2 , 1 6 2 , 5 0 2 , 5 6 2 機械学習部
- 1 0 3 , 1 6 3 , 5 0 3 , 5 6 3 学習結果データベース
- 1 1 0 , 5 0 4 素性抽出部
- 1 1 1 , 1 7 1 , 5 0 5 , 5 6 5 解推定処理部
- 1 6 1 , 5 6 1 素性 - 解対・素性 - 解候補対抽出部
- 1 7 0 , 5 6 4 素性 - 解候補抽出部
- 2 0 1 問題表現相当部抽出部
- 2 0 2 問題表現情報記憶部
- 2 0 3 意味解析情報記憶部
- 2 0 4 問題構造変換部
- 2 0 5 教師なしデータ記憶部
- 5 0 0 , 5 4 0 , 5 5 0 , 5 8 0 , 6 0 0 , 6 5 0 言語解析処理システム
- 5 1 1 , 5 2 1 , 6 1 1 , 6 2 1 第 1 素性追加部
- 5 1 2 , 5 2 2 , 6 1 2 , 6 2 2 第 2 素性追加部
- 1 0 1 0 , 1 0 3 0 スタック用教師なし学習処理システム
- 1 0 2 0 , 1 0 4 0 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

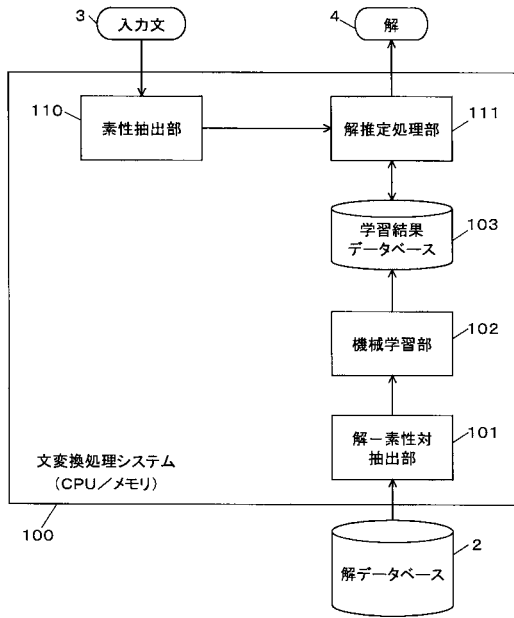
10

20

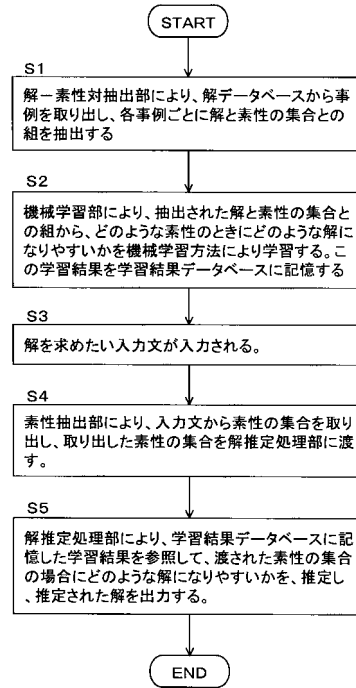
30

40

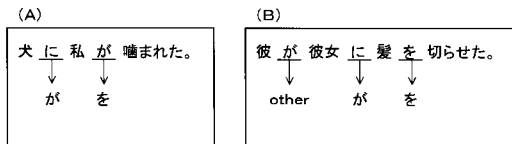
【 図 1 】



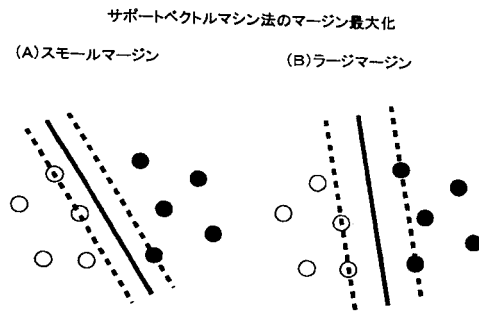
【 図 2 】



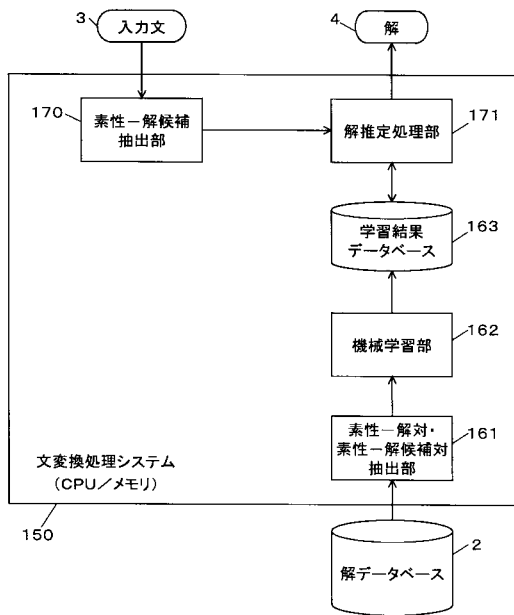
【 図 3 】



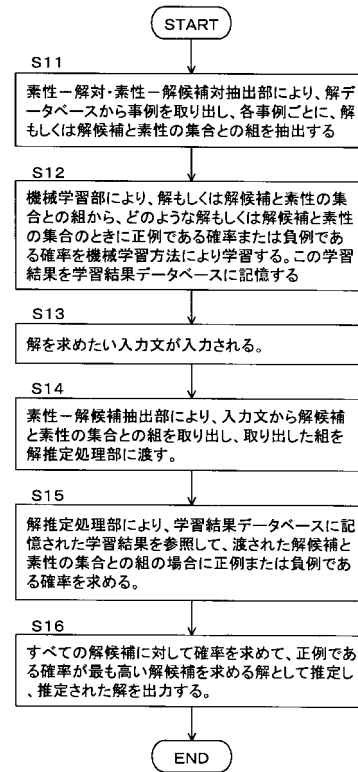
【 図 4 】



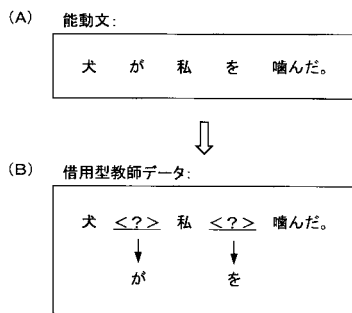
【 図 5 】



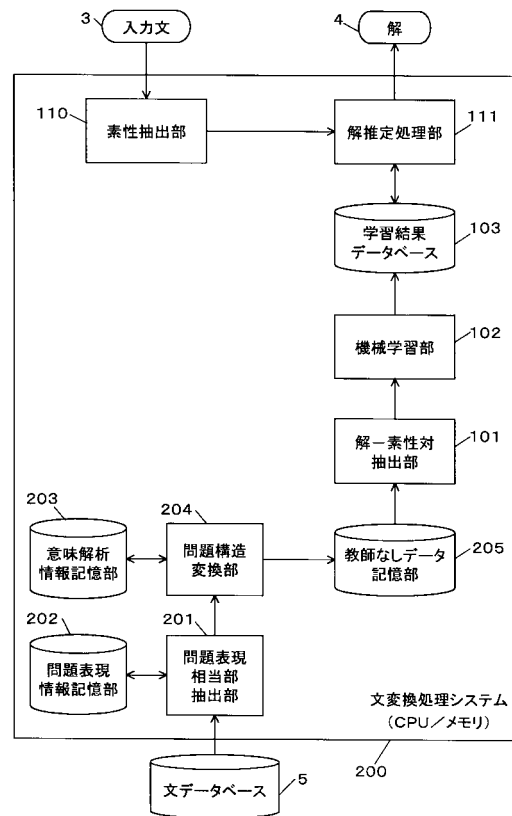
【 図 6 】



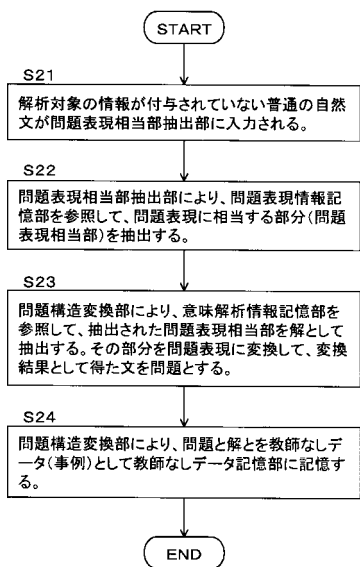
【 図 7 】



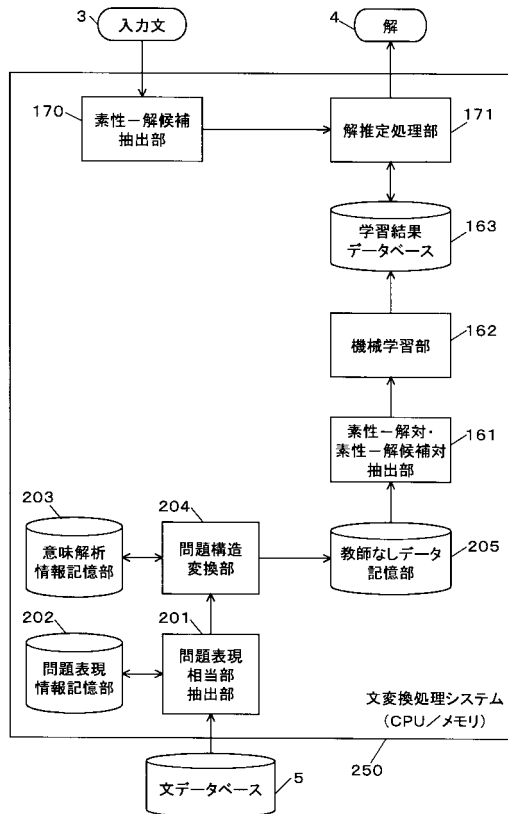
【 図 8 】



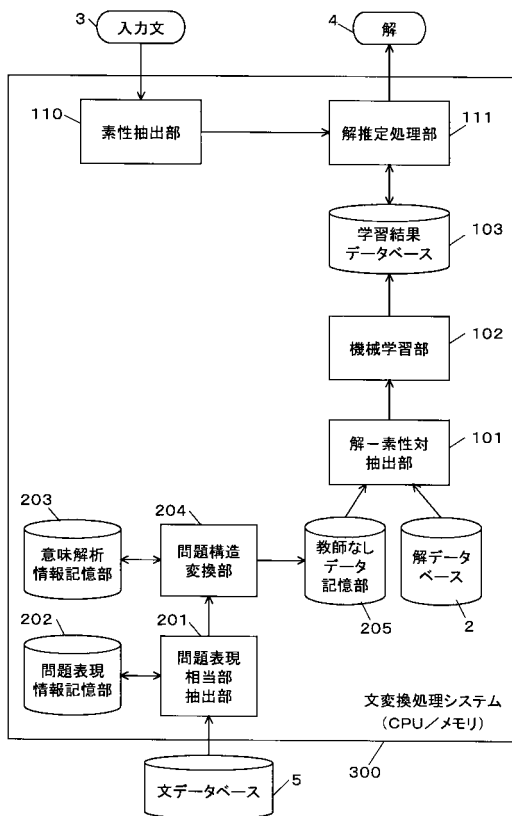
【図9】



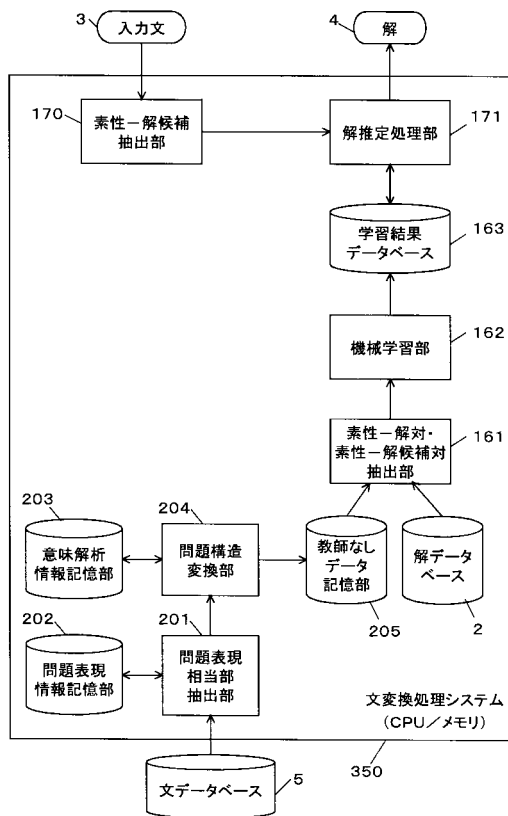
【図10】



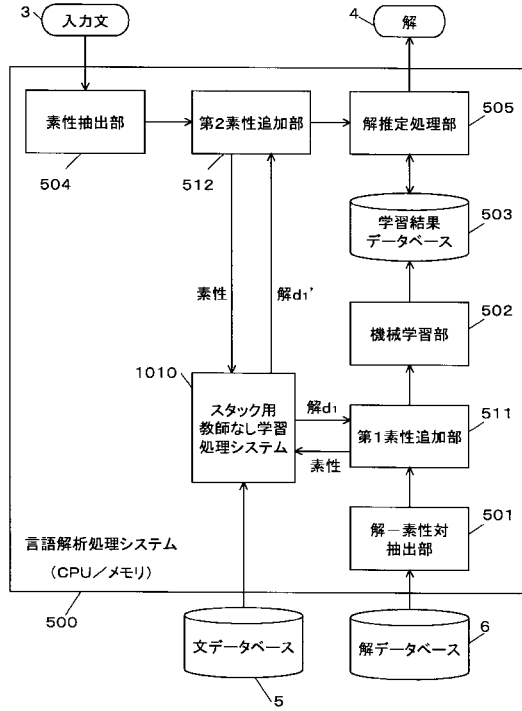
【図11】



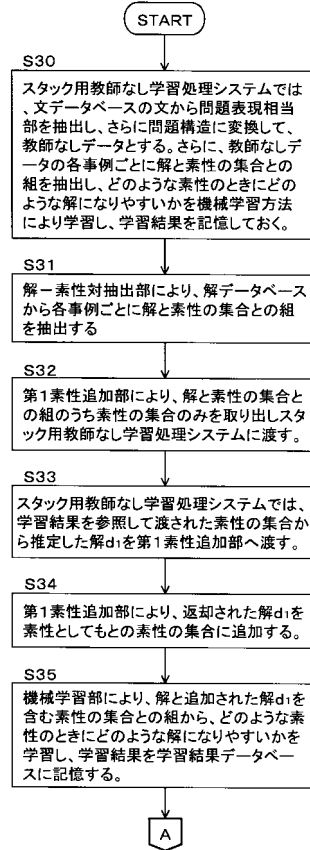
【図12】



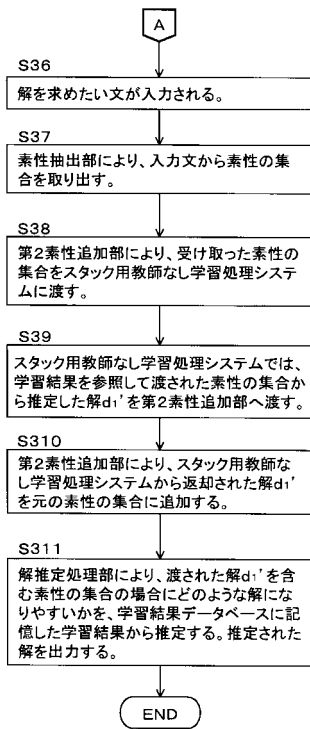
【図13】



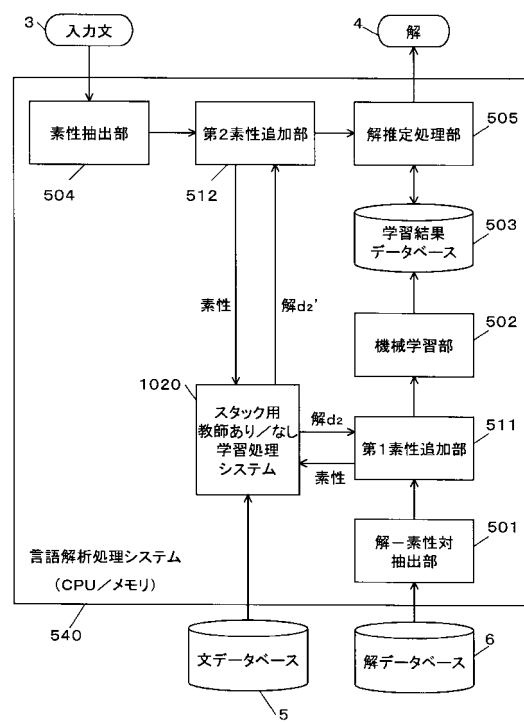
【図14】



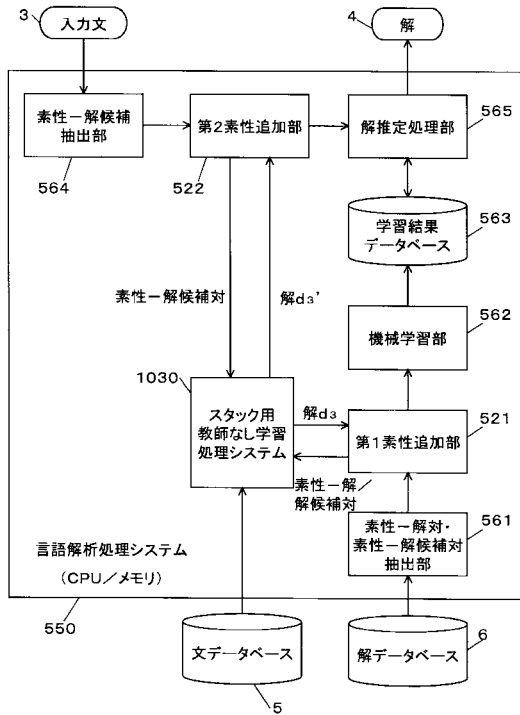
【図15】



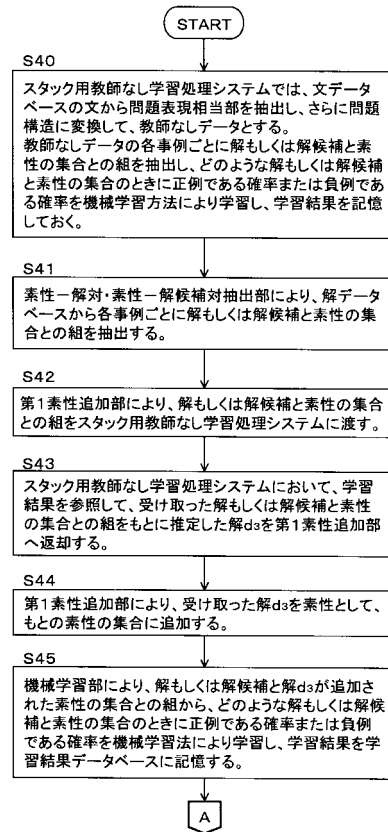
【図16】



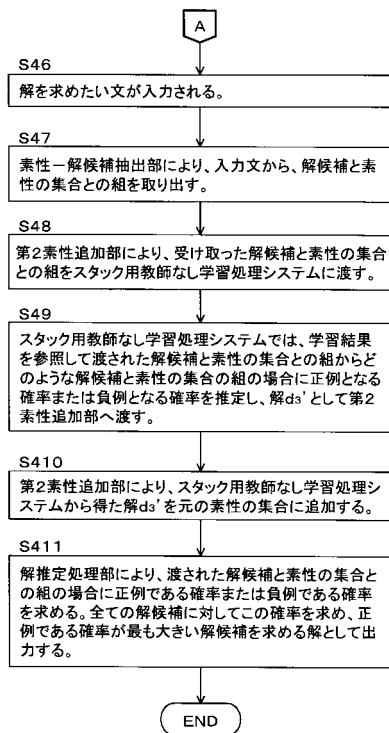
【 図 1 7 】



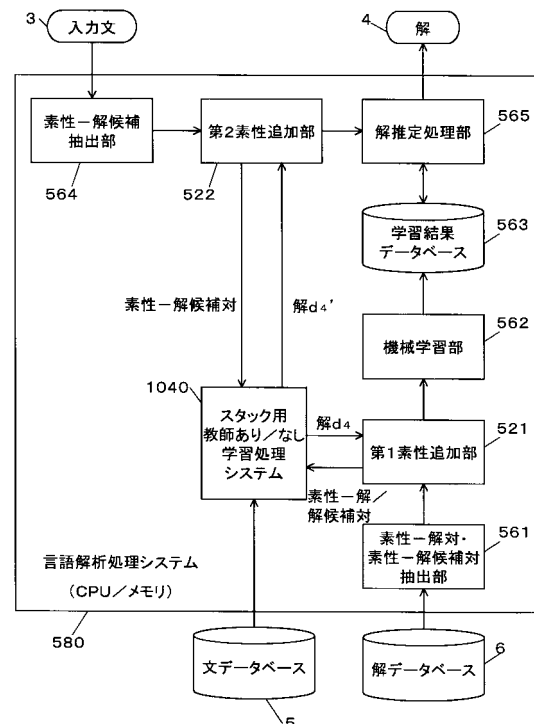
【 図 1 8 】



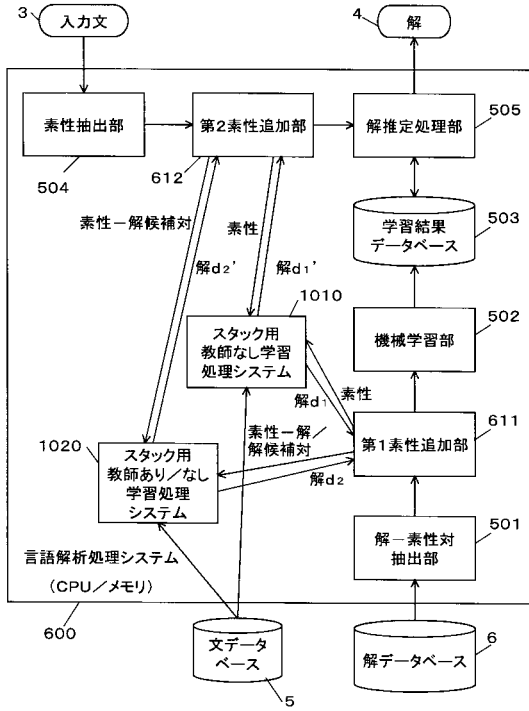
【 図 1 9 】



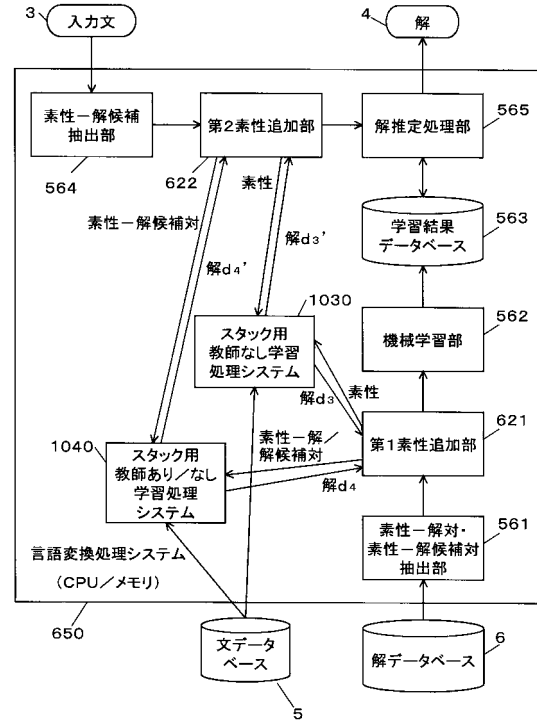
【 図 2 0 】



【図 2 1】



【図 2 2】



【図 2 3】

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