(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第3899414号
(P3899414)
(45)発行日　平成19年3月28日(2007.3.28)　(24)登録日　平成19年1月12日(2007.1.12)

(51)Int.Cl.　F I
G06F 17/28 (2006.01)　G06F 17/28 U

請求項の数 15　(全 34 頁)

(21)出願番号　特願2004-103862(P2004-103862)
(22)出願日　平成16年3月31日(2004.3.31)
(65)公開番号　特開2005-292958(P2005-292958A)
(43)公開日　平成17年10月20日(2005.10.20)
審査請求日　平成16年3月31日(2004.3.31)

(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

審査官　和田　財太

(56)参考文献　特開２００３－１２２７５０（ＪＰ，Ａ）
特開２００３－２４８６７６（ＪＰ，Ａ）

最終頁に続く

(54)【発明の名称】教師データ作成装置およびプログラム、ならびに言語解析処理装置およびプログラム

(57)【特許請求の範囲】

【請求項１】

機械学習法を用いた所定の言語解析処理において使用する教師データをコーパスから作成する教師データ作成装置であって、

テキストデータで構成されるコーパスを入力するコーパス入力手段と、

前記コーパスのテキストデータにおいて、ユーザによって選択された文字列の前後に、所定の言語解析の結果となる言語情報の個々の分類先を示すタグであってマークアップ言語の形式で記述された分類タグを挿入する分類タグ付与手段と、

前記分類タグが挿入されたテキストデータにおいて、前記分類タグが挿入された箇所を含む所定の範囲の前後に、教師データとして使用する範囲を示すタグであってマークアップ言語の形式で記述された範囲指定タグを挿入する範囲指定タグ付与手段と、

前記分類タグおよび前記範囲指定タグが挿入されたテキストデータから、前記範囲指定タグに囲まれたデータをユーザ範囲データとして抽出するユーザ範囲抽出手段とを、備える

ことを特徴とする教師データ作成装置。

【請求項２】

ユーザによって入力された分類タグをタグ記憶手段に記憶するタグ登録手段を備え、

前記分類タグ付与手段は、前記タグ記憶手段に記憶された分類タグを前記コーパスのテキストデータに挿入する

ことを特徴とする請求項１記載の教師データ作成装置。

【請求項３】
前記範囲指定タグ付与手段は、前記分類タグが挿入されたテキストデータにおいて、ユーザによって指定された前記分類タグが挿入された箇所を含む範囲の前後に前記範囲指定タグを挿入する
ことを特徴とする請求項１記載の教師データ作成装置。
【請求項４】
前記範囲指定タグ付与手段は、前記分類タグが挿入されたテキストデータにおいて、前記分類タグが挿入された箇所を含む所定の範囲を所定の範囲指定規則にもとづいて指定し、前記指定された範囲の前後に前記範囲指定タグを挿入する
ことを特徴とする請求項１記載の教師データ作成装置。
【請求項５】
ユーザによって定義されたユーザ範囲指定規則を規則記憶手段に記憶する規則登録手段を備え、
前記範囲指定タグ付与手段は、前記規則記憶手段に記憶されたユーザ範囲指定規則に従って前記範囲指定タグを挿入する
ことを特徴とする請求項４記載の教師データ作成装置。
【請求項６】
前記分類タグ付与手段は、前記テキストデータにおいて、前記分類タグが挿入された箇所を含む所定の範囲の前後に、前記分類タグのうちユーザによって指定された特定の分類先だけに対する教師データとして使用する範囲を示すタグであってマークアップ言語の形式で記述されたユーザ指定分類タグ用範囲指定タグを付与し、
前記ユーザ範囲抽出手段は、前記ユーザ指定分類タグ用範囲指定タグが挿入されたテキストデータから、前記ユーザ指定分類タグ用範囲指定タグに囲まれたデータを、前記特定の分類先に対する教師データを生成するためのユーザ範囲データとして抽出する
ことを特徴とする請求項１記載の教師データ作成装置。
【請求項７】
さらに、前記ユーザ範囲データを所定の単位ごとに分割し、前記ユーザ範囲データから前記分類タグに囲まれた文字列を検出し、前記分割した単位のうち前記検出した文字列に対応する部分に前記分類タグに対応する分類先を前記単位ごとに付与し、各単位のデータを、解を前記分類先とする教師データに変換する教師データ変換手段を備える
ことを特徴とする請求項１記載の教師データ作成装置。
【請求項８】
前記教師データ変換手段は、前記検出した文字列が複数の単位である場合に、前記分類先に前記文字列における単位の位置を示す情報を付加したものを、単位ごとに付与する
ことを特徴とする請求項７記載の教師データ作成装置。
【請求項９】
さらに、前記教師データから所定の種類の素性を抽出し、前記単位について、前記素性の集合と前記付与された分類先との組を生成する素性抽出手段を備える
ことを特徴とする請求項７記載の教師データ作成装置。
【請求項１０】
前記素性抽出手段は、前記教師データに対して形態素解析を行い所定の種類の素性を抽出する
ことを特徴とする請求項９記載の教師データ作成装置。
【請求項１１】
前記素性抽出手段は、前記教師データから所定の文字または文字列を切り出して素性とする
ことを特徴とする請求項９記載の教師データ作成装置。
【請求項１２】
教師データを用いた機械学習法により所定の言語解析処理を行う言語解析処理装置であって、

テキストデータで構成されるコーパスであって、所定の言語解析の結果となる言語情報の個々の分類先を示すタグであってマークアップ言語の形式で記述された分類タグと、前記分類タグが挿入された箇所を含む所定の範囲の前後に、教師データとして使用する範囲を示すタグであってマークアップ言語の形式で記述された範囲指定タグとが付与されたものを入力し、前記コーパスから、前記範囲指定タグに囲まれたデータをユーザ範囲データとして抽出するユーザ範囲抽出手段と、

前記ユーザ範囲データを所定の単位ごとに切り出し、前記ユーザ範囲データから前記分類タグに囲まれた文字列を検出し、前記切り出した単位のうち前記検出した文字列に対応するものに前記分類タグに対応する分類先を付与し、前記切り出した単位のうち前記検出した文字列に対応しないものに分類先がないことを示す分類先を付与し、単位ごとのデータを教師データとする教師データ変換手段と、

前記教師データから所定の種類の素性を抽出し、前記単位について、前記素性の集合と前記付与された分類先との組を生成する素性抽出手段と、

前記素性の集合と前記分類先との組を利用して、前記単位について、前記素性の集合の場合にどのような分類先になりやすいかを学習し、前記学習の結果を記憶しておく機械学習手段と、

言語解析処理の対象とするテキストデータを入力するデータ入力手段と、

前記入力データから所定の解析処理または切り出し処理により素性を抽出する所定の種類の素性を抽出する素性抽出手段と、

前記学習結果を利用して、前記入力データの所定の単位のデータについて、前記素性の場合になりやすい分類先を推定する解推定手段と、

前記推定された分類先に対応する分類タグを、前記入力データの前記推定の対象となった単位に対応する文字列の前後に挿入するタグ付与手段とを、備える

ことを特徴とする言語解析処理装置。

【請求項１３】

さらに、分類タグが挿入された前記入力データから、前記分類タグに囲まれた文字列を、前記分類タグに囲まれていない文字列と異なる表示態様で表示する解析結果表示処理手段を備える

ことを特徴とする請求項１２記載の言語解析処理装置。

【請求項１４】

請求項１記載の教師データ生成装置として、コンピュータを機能させるための教師データ生成プログラム。

【請求項１５】

請求項１２記載の言語解析処理装置として、コンピュータを機能させるための言語解析処理プログラム。

【発明の詳細な説明】

【技術分野】

【０００１】

本発明は、言語データの解析処理で実行される機械学習処理用の教師データをコーパスから作成する教師データ作成装置に関する。また、本発明は、前記教師データ作成装置によって作成された教師データを用いて機械学習法による言語データの解析処理を行う言語解析処理装置もしくは要約処理装置、および前記処理装置としてコンピュータを機能させるためのプログラムに関する。

【０００２】

コーパスとは、コンピュータが読み取り可能な大量のテキストデータなどの言語資源であって、例えば新聞記事、小説、論文などの文章の電子化データである。

【背景技術】

【０００３】

機械学習法を用いた言語データの解析処理では、精度の良い機械学習を実現するために教師あり機械学習法が採用されることが多い。教師あり機械学習では、学習処理過程の教

師データとして、テキストデータに解析処理の結果となるような言語情報、例えば品詞情報、文字種情報、照応関係情報、意味情報などが付与された加工済みコーパスが利用される。

【０００４】

そして、言語情報が付与された加工済みコーパスを、言語情報が付与されていないテキストデータだけで構成される生コーパスから生成する場合に、原則として、コーパスの全てのテキストデータに対して言語情報を付与しておくことが必要である。一部分のテキストデータにのみ言語情報が付与されているようなコーパスを教師データとして用いて機械学習を行った場合には、機械学習の学習精度が低下するからである。

【０００５】

例えば、機械学習により人名や地名などの固有表現を抽出する処理（固有表現抽出処理）をする場合に教師データとするコーパスを想定する。付与される言語情報は、その名詞がどのような固有表現の分類であるかを示す分類ラベル（人名、地名など）である。分類ラベルを付与するためのコーパスのテキストデータの一部に、以下のような部分があるとする。

「…日本の首相は小泉さんです。小泉さんはいつも思いきったことをしています。…」

ユーザは、コーパス中の文「日本の首相は小泉さんです。」だけをチェックし、文中の単語「日本」に分類ラベル「地名」を、単語「小泉」に分類ラベル「人名」を付与する作業をしたとする。作業後の文は、以下のような状態になる。

「…”日本（地名）”の首相は”小泉（人名）”さんです。小泉さんはいつも思いきったことをしています。…」

このような言語情報（分類ラベル）を部分的にのみ付与したコーパスを教師データとして機械学習し、その学習結果を用いて固有表現抽出処理を行うとする。学習処理段階では、コーパス内の個々の単語の所定の素性を抽出し、付与された分類ラベルをもとに、その単語が「どのような素性の場合にどのような分類先になりやすいか」を学習する。素性とは、所定の解析処理のために用いる情報（例えば、品詞情報、字種情報、係り受け関係のような統語情報など）の一単位であって、文字や形態素などの所定の単位が備える性質を意味する。

【０００６】

ここで、単語ごとの各分類先へのなりやすさを評価する場合に、単語「小泉」の分類先「人名」へのなりやすさは、最初の文「”日本（地名）”の首相は”小泉（人名）”さんです。」の単語「小泉」に付与された分類ラベル「人名」によって高いスコアとなる。しかし、２番目の文「小泉さんはいつも思いきったことをしています。」の単語「小泉」には分類ラベル「人名」が付与されていないため、２番目の文中の単語「小泉」によって、「小泉」の分類先「人名」へのなりやすさのスコアは低下する。この２番目の文中の単語「小泉」のように、単にユーザが言語情報の付与作業をしなかった単語が存在することにより、学習処理での評価精度が低下してしまうことは問題である。

【０００７】

したがって、コーパスの全体に所定の言語情報を付与する必要があるが、付与作業は、作業量が非常に膨大であり処理負担が大きい。そのため、通常、コーパスを利用した教師データは、言語解析処理装置の管理者や開発者によって準備されることが多い。ユーザの教師データ作成作業の負担を軽減するために、生コーパスから教師データを自動的に生成して機械学習で利用する技術がある（例えば、特許文献１参照）。

【０００８】

また、言語データ解析処理の一つとして、文章データからその内容を表わすために重要と考えられる文（重要文という）を抽出して自動的に要約を生成する要約処理装置がある。要約結果に対する評価はユーザ個人の指向や要約の用途によって相違することが知られている。そのため、個々のユーザの指向や用途に適応した要約処理が行えるようにする必要がある。そのため、文章データに対する要約に対するユーザの評価を機械学習法を用いて学習し、ユーザに適応した要約処理を行えるようにする技術がある（例えば、特許文献

２参照）。
【特許文献１】特開２００３－１２２７５０
【特許文献２】特開２００３－２４８６７６
【発明の開示】
【発明が解決しようとする課題】
【０００９】
特許文献１の技術のように、生コーパスから教師データを作成する手法として、生コーパスの広範かつ多数のデータから人手によらずに自動的に教師データを作成し、豊富な教師データによって機械学習の学習精度の向上を目指す手法がある。
【００１０】
しかし、生コーパスの多量なデータの部分について言語情報を人手により確実に付与することにより、正確な言語情報が付与された教師データによって機械学習の学習精度の向上を目指すことも可能である。この場合、多量のデータで構成される生コーパスを部分的に使用することによって生ずる学習精度の低下を防止する必要がある。
【００１１】
また、生コーパスで部分的に言語情報を付与する範囲を、ユーザが自由に選択でき、かつ、選択した範囲を自由に追加できれば、システムの管理者や開発者だけでなく一般的なユーザも教師データを作成することができ、開発の負担を軽減することができる。
【００１２】
また、ユーザが任意に定義した言語情報を簡単に付与できれば、さまざまな言語情報を解析対象とすることができる。
【００１３】
また、要約処理において、ユーザが要約に重要と考える文（重要文）をユーザが簡単に指示できれば、ユーザの指向に適合した要約処理のための教師データの作成処理負担を軽減することができる。
【００１４】
本発明の目的は、機械学習法を用いた言語データの解析処理において使用する教師データをコーパスから作成する場合に、解析処理の結果となる言語情報、言語情報を付与する文字、文節、単語などの箇所を、ユーザがインタラクティブなインタフェースにより自由かつ簡単に指定でき、大量なデータで構成されるコーパスの一部にのみ言語情報の付与を行った場合でも、言語情報の付与作業が確認された範囲を特定して教師データを作成できるような教師データ作成装置を提供することである。
【００１５】
また、本発明の目的は、機械学習法を用いた言語データの解析処理を行う場合に、言語情報が一部のデータにのみ付与されているようなコーパスから、ユーザによる言語情報の付与が確認された範囲のデータのみを教師データとして使用し、学習精度を低下させずに機械学習を行えるような言語解析処理装置を提供することである。
【００１６】
また、本発明の目的は、機械学習法を用いた要約処理を行う場合に、ユーザの指向に適応した要約を学習し、文章の要約を行えるような要約処理装置を提供することである。
【００１７】
また、本発明の目的は、前記処理装置としてコンピュータを機能させるためのプログラムを提供することである。
【課題を解決するための手段】
【００１８】
本発明は、機械学習法を用いた所定の言語データの解析処理において使用する教師データをコーパスから作成する教師データ作成装置であって、１）テキストデータで構成されるコーパスを入力するコーパス入力手段と、２）前記コーパスのテキストデータにおいて、ユーザによって選択された文字列の前後に、所定の言語解析の結果となる言語情報の個々の分類先を示すタグであってマークアップ言語の形式で記述された分類タグを挿入する

分類タグ付与手段と、３）前記分類タグが挿入されたテキストデータにおいて、前記分類タグが挿入された箇所を含む所定の範囲の前後に、教師データとして使用する範囲を示すタグであってマークアップ言語の形式で記述された範囲指定タグを挿入する範囲指定タグ付与手段と、４）前記分類タグおよび前記範囲指定タグが挿入されたテキストデータから、前記範囲指定タグに囲まれたデータをユーザ範囲データとして抽出するユーザ範囲抽出手段とを、備える。

【００１９】

本発明の、分類タグ付与手段では、所定の言語データの解析処理において、結果となる言語情報の個々の分類先を示す分類タグを用意しておく。

【００２０】

また、範囲指定タグ付与手段では、分類タグが挿入された箇所を含む所定の範囲であって、教師データとして使用される範囲を示す範囲指定タグを用意しておく。

【００２１】

分類タグおよび範囲指定タグは、ＳＧＭＬ（Ｓｔａｎｄａｒｄ Ｇｅｎｅｒａｌｉｚｅｄ Ｍａｒｋｕｐ Ｌａｎｇｕａｇｅ）形式で記述される属性情報であって、指定された箇所（文字、単語、文節、文など）を挟むようにタグが挿入されることによって、タグで囲まれた部分に付与される所定の言語情報（分類先）を表現するものである。

【００２２】

そして、コーパス入力手段が、テキストデータで構成されるコーパスを入力すると、分類タグ付与手段では、前記コーパスのテキストデータにおいて、ユーザによって選択された文字列の前後に分類タグを挿入し、範囲指定タグ付与手段では、前記分類タグが挿入されたテキストデータにおいて、前記分類タグが挿入された箇所を含む所定の範囲の前後に、範囲指定タグを挿入する。そして、ユーザ範囲抽出手段では、前記分類タグおよび前記範囲指定タグが挿入されたテキストデータから、前記範囲指定タグに囲まれたデータを、教師データを生成するためのユーザ範囲データとして抽出する。

【００２３】

本発明の教師データ作成装置では、ユーザが、教師データを作成するために、膨大なデータ量のコーパスの中から必要な範囲のデータにだけ言語情報を付与するような作業を行った場合でも、ユーザがチェックした範囲のデータだけを、教師データ作成のために抽出することができる。これにより、従来のように同一単語に言語情報が付与されていたり付与されていなかったりする状態に因る機械学習の学習精度の低下を生じさせないような教師データを作成することができる。

【００２４】

また、本発明の教師データ作成装置は、上記構成をとる場合に、さらに、ユーザによって入力された分類タグをタグ記憶手段に記憶するタグ登録手段を備え、前記分類タグ付与手段は、前記タグ記憶手段に記憶された分類タグを前記コーパスのテキストデータに挿入するものである。これにより、教師データにおいて解となる分類先を、ユーザが任意に設定することができる。

【００２５】

また、本発明の教師データ作成装置は、前記範囲指定タグ付与手段が、前記分類タグが挿入されたテキストデータにおいて、ユーザによって指定された前記分類タグが挿入された箇所を含む範囲の前後に前記範囲指定タグを挿入するものである。これにより、教師データ作成のために、コーパスからユーザが分類タグの付与をチェックした範囲のデータだけを抽出することができる。

【００２６】

また、本発明の教師データ作成装置は、前記範囲指定タグ付与手段が、前記分類タグが挿入されたテキストデータにおいて、前記分類タグが挿入された箇所を含む所定の範囲を所定の範囲指定規則にもとづいて指定し、前記指定された範囲の前後に前記範囲指定タグを挿入するものである。これにより、教師データ作成のために、ユーザが分類タグを指定するだけで、コーパスからユーザが分類先の付与をチェックした範囲のデータだけを抽出

することができる。

【００２７】

また、本発明の教師データ作成装置は、さらに、前記ユーザ範囲データを所定の単位ごとに分割し、前記ユーザ範囲データから前記分類タグに囲まれた文字列を検出し、前記分割した単位のうち前記検出した文字列に対応する部分に前記分類タグに対応する分類先を前記単位ごとに付与し、各単位のデータを、解を前記分類先とする教師データに変換する教師データ変換手段を備える。

【００２８】

これにより、コーパスから、ユーザが分類先をチェックした範囲のデータを用いて、ユーザが指定した分類先を解とする教師データを作成することができる。

【００２９】

さらに、本発明は、教師データを用いた機械学習法により所定の言語解析処理を行う言語解析処理装置であって、１）テキストデータで構成されるコーパスであって、所定の言語解析の結果となる言語情報の個々の分類先を示すタグであってマークアップ言語の形式で記述された分類タグと、前記分類タグが挿入された箇所を含む所定の範囲の前後に、教師データとして使用する範囲を示すタグであってマークアップ言語の形式で記述された範囲指定タグとが付与されたものを入力し、前記コーパスから、前記範囲指定タグに囲まれたデータをユーザ範囲データとして抽出するユーザ範囲抽出手段と、２）前記ユーザ範囲データを所定の単位ごとに切り出し、前記ユーザ範囲データから前記分類タグに囲まれた文字列を検出し、前記切り出した単位のうち前記検出した文字列に対応するものに前記分類タグに対応する分類先を付与し、前記切り出した単位のうち前記検出した文字列に対応しないものに分類先がないことを示す分類先を付与し、単位ごとのデータを教師データとする教師データ変換手段と、３）前記教師データから所定の種類の素性を抽出し、前記単位について、前記素性の集合と前記付与された分類先との組を生成する素性抽出手段と、４）前記素性の集合と前記分類先との組を利用して、前記単位について、前記素性の集合の場合にどのような分類先になりやすいかを学習し、前記学習の結果を記憶しておく機械学習手段と、５）言語解析処理の対象とするテキストデータを入力するデータ入力手段と、６）前記入力データから所定の解析処理または切り出し処理により素性を抽出する所定の種類の素性を抽出する素性抽出手段と、７）前記学習結果を利用して、前記入力データの所定の単位のデータについて、前記素性の場合になりやすい分類先を推定する解推定手段と、８）前記推定された分類先に対応する分類タグを、前記入力データの前記推定の対象となった単位に対応する文字列の前後に挿入するタグ付与手段とを、備える。

【００３０】

本発明の言語解析処理装置は、コーパスから、ユーザが分類先をチェックした範囲のデータを用いて、ユーザが指定した分類先を解とする教師データを作成し、この教師データを利用した機械学習法により所定の言語解析処理を行う。

【００３１】

本発明の言語解析処理装置は、ユーザ範囲抽出手段により、前記教師データ作成装置により作成された、テキストデータで構成されるコーパスであって、所定の言語解析の結果となる言語情報の個々の分類先を示すタグであってマークアップ言語の形式で記述された分類タグと、前記分類タグが挿入された箇所を含む所定の範囲の前後に、教師データとして使用する範囲を示すタグであってマークアップ言語の形式で記述された範囲指定タグとが付与されたものを入力し、入力したコーパスから、前記範囲指定タグに囲まれたデータを、教師データを生成するためのユーザ範囲データとして抽出する。そして、教師データ変換手段では、前記ユーザ範囲データを所定の単位（形態素、文字、文字列、単語、文節、文など）ごとに切り出し、前記ユーザ範囲データから前記分類タグに囲まれた文字列を検出し、前記切り出した単位のうち前記検出した文字列に対応するものに前記分類タグに対応する分類先を付与し、前記切り出した単位のうち前記検出した文字列に対応しないものに分類先がないことを示す分類先を付与し、単位ごとのデータを教師データとする。

【００３２】

さらに、素性抽出手段は、前記教師データから所定の種類の素性を抽出し、前記単位について、前記素性の集合と前記付与された分類先との組を生成する。そして、機械学習手段では、前記素性の集合と前記分類先との組を利用して、前記単位について、前記素性の集合の場合にどのような分類先になりやすいかを学習し、前記学習の結果を記憶しておく。

【００３３】

その後、データ入力手段により、言語解析処理の対象とするテキストデータを入力すると、素性抽出手段では、前記入力データから所定の解析処理または切り出し処理により素性を抽出し、解推定手段では、前記学習結果を利用して、前記入力データの所定の単位のデータについて、前記素性の場合になりやすい分類先を推定する。そして、タグ付与手段では、前記推定された分類先に対応する分類タグを、前記入力データの前記推定の対象となった単位に対応する文字列の前後に挿入する。

【００３４】

これにより、ユーザが、膨大なデータ量のコーパスの中から必要な範囲のデータにだけ言語情報（分類先）を付与して言語解析処理を行うような場合でも、従来のように同一単語に言語情報が付与されていたり付与されていなかったりする状態に因る機械学習の学習精度の低下を生じさせないような機械学習による言語解析処理を行うことができる。

【００３５】

また、本発明は、教師データを用いた機械学習法により文章の要約を行う要約処理装置であって、１）複数の文で構成される教師用のテキストデータを入力する教師用データ入力手段と、２）前記テキストデータにおいて、ユーザによって選択された文の前後に、要約処理において重要な文であることを示すタグであってマークアップ言語の形式で記述された重要文タグを挿入する重要文タグ付与手段と、３）前記重要文タグが挿入されたテキストデータにおいて、前記重要文タグが挿入された文が含まれる要約する対象となる文章の範囲の前後に、教師データとして使用する範囲を示すタグであってマークアップ言語の形式で記述された範囲指定タグを挿入する範囲指定タグ付与手段と、４）前記重要文タグおよび前記範囲指定タグが挿入されたテキストデータから、前記範囲指定タグに囲まれたデータをユーザ範囲データとして抽出するユーザ範囲抽出手段と、５）前記ユーザ範囲データを文単位に分割し、前記ユーザ範囲データから前記重要文タグに囲まれた文を検出し、前記分割した文のうち前記検出した文に重要文であることを示す分類先を付与し、前記分割した文のうち前記検出した文以外の文に重要文でないことを示す分類先を付与し、各文を教師データとする教師データ変換手段と、６）前記教師データから所定の種類の素性を抽出し、前記文について、前記素性の集合と前記付与された分類先との組を生成する素性抽出手段と、７）前記文についての前記素性と前記分類先との組を利用して、前記各文について、前記素性の集合の場合にどのような分類先になりやすいかを学習し、前記学習の結果を記憶しておく機械学習手段と、８）要約の対象とするテキストデータを入力するデータ入力手段と、９）前記入力データから所定の解析処理または切り出し処理により所定の種類の素性を抽出する素性抽出手段と、１０）前記学習結果を利用して、前記入力データの各文について、前記素性の場合になりやすい分類先を推定する解推定手段と、１１）前記推定された分類先が重要文である文の前後に重要文タグを挿入するタグ付与手段と、１２）前記入力データの前記重要文タグで囲まれた文を要約として出力する要約出力処理手段とを、備える。

【００３６】

本発明の要約処理装置は、教師データ用のテキストデータにおいて、要約上重要な文であるとしてユーザが指定した文を含む範囲のデータをもとに教師データを作成し、この教師データを利用した機械学習法により、要約対象のテキストデータの要約処理を行う。

【００３７】

本発明の要約処理装置は、教師用データ入力手段により、複数の文で構成される教師用のテキストデータを入力すると、重要文タグ付与手段では、前記教師用のテキストデータにおいて、ユーザによって選択された文の前後に、要約処理において重要な文であること

を示すタグであってマークアップ言語の形式で記述された重要文タグを挿入し、範囲指定タグ付与手段では、前記重要文タグが挿入されたテキストデータにおいて、前記重要文タグが挿入された文が含まれる要約の対象となる文章の範囲の前後に、範囲指定タグを挿入する。そして、ユーザ範囲抽出手段では、前記重要文タグおよび前記範囲指定タグが挿入されたテキストデータから、前記範囲指定タグに囲まれたデータをユーザ範囲データとして抽出する。

【００３８】

そして、教師データ変換手段では、前記ユーザ範囲データを文単位に分割し、前記ユーザ範囲データから前記重要文タグに囲まれた文を検出し、前記分割した文のうち前記検出した文に重要文であることを示す分類先を付与し、前記分割した文のうち前記検出した文以外の文に重要文でないことを示す分類先を付与し、各文を教師データとする。

【００３９】

さらに、素性抽出手段では、前記教師データから所定の種類の素性を抽出し、前記文について、前記素性の集合と前記付与された分類先との組を生成する。そして、機械学習手段では、前記文についての前記素性と前記分類先との組を利用して、前記各文について、前記素性の集合の場合にどのような分類先になりやすいかを学習し、前記学習の結果を記憶しておく。

【００４０】

その後、データ入力手段により、要約対象のテキストデータを入力すると、素性抽出手段では、前記入力データから所定の解析処理または切り出し処理により所定の種類の素性を抽出する。解推定手段では、前記学習結果を利用して、前記入力データの各文について、前記素性の場合になりやすい分類先を推定する。そして、タグ付与手段では、前記推定された分類先が重要文である文の前後に重要文タグを挿入し、要約出力処理手段では、前記入力データの前記重要文タグで囲まれた文を要約として出力する。

【００４１】

これにより、ユーザが、自分の嗜好や要約の用途などに応じて重要文を指定することができ、ユーザが選択した重要文による機械学習によりユーザ各々に適応した要約を作成することができる。

【００４２】

なお、本発明は、本発明の教師データ作成装置、言語解析処理装置、または要約処理装置としてコンピュータを機能させるためのプログラムとして実現することができる。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。

【発明の効果】

【００４３】

本発明によれば、機械学習法を用いた言語データの解析処理において使用する教師データをコーパスから作成する場合に、解析処理の結果となる言語情報、言語情報を付与する文字、文節、単語などの箇所を、ユーザがインタラクティブなインタフェースにより自由かつ簡単に指定でき、コーパスの言語情報の付与作業が確認された範囲を特定できる。

【００４４】

これにより、ユーザは任意に定義した言語情報をコーパスのような多量なテキストデータの任意な箇所に付与して教師データを作成することができ、あるコーパスを用いて徐々に教師データを増加させていくような作業を可能とするため、過度の作業負担を軽減することができる。

【００４５】

また、本発明によれば、機械学習法を用いた言語データの解析処理を行う場合に、言語情報が一部のデータにのみ付与されているようなコーパスから、ユーザによる言語情報の付与が確認された範囲のデータのみを教師データとして使用し、学習精度を低下させずに

機械学習を行うことができる。これにより、言語情報の付与作業が途中であるようなコーパスを教師データとして使用することができる。また、部分的にタグが付与されたようなコーパスを効率的に利用することができる。

【００４６】

特に、教師データ作成の専門家ではないようなユーザが、コーパスに分類先などの言語情報をタグ付けする場合に、膨大なデータ量のコーパスのすべてにタグ付け作業を行うことは困難であり、コーパスの部分部分に対してのみタグ付け作業を行うことが予想される。このような状態でタグ付けがなされたコーパスからでも、機械学習の処理精度を低下させない教師データを作成することができる。

【００４７】

また、本発明によれば、一般的なユーザが機械学習法を用いた処理装置を利用したい場合に、コーパスに大規模なタグ付け作業を行うことなく、部分的に言語情報のタグを付与するだけでよいため、手軽に機械学習法を用いた処理装置を利用できるようになる。

【００４８】

さらに、本発明によれば、ユーザが任意に定義した分類タグを付与することができる。すなわち、本発明では、ユーザ自身が興味を持った問題を解いたり、興味を持った表現を抽出するために分類タグを定義し、ユーザ自身で簡単にコーパスに付与することができる。さらに、このようなユーザの興味にもとづく分類タグを付与された教師データを利用した機械学習を行うことにより、言語解析処理装置は、ユーザが興味を持つ表現などを抽出することができるようになる。その結果、ユーザは、機械学習法を用いた言語解析処理装置を自身の知的活動の一部として利用することが可能になる。

【００４９】

また、本発明によれば、機械学習法を用いた要約処理を行う場合に、ユーザは、要約として重要と考えるような文（重要文）を自由かつ簡単に指定して、文章の要約を行うことができる。これにより評価が分かれやすい要約処理について、各ユーザに適応した要約を出力することができる。

【発明を実施するための最良の形態】

【００５０】

以下、図を用いて本発明を実施するための最良の形態を説明する。

【００５１】

図１は、機械学習法を用いた言語解析処理を行う場合の本発明の構成例を示す図である。

【００５２】

教師データ作成装置１は、ＣＰＵおよびメモリを備えて、機械学習法を用いた言語解析処理で使用する教師データを作成する装置であって、コーパス入力手段１１、タグ登録手段１２、タグ記憶手段１３、タグ付与手段１４、コーパス記憶手段１５、ユーザ範囲抽出手段１６、教師データ変換手段１７、規則登録手段１８、規則記憶手段１９、素性抽出手段１１０、表示装置２１、および入力装置２２を備える。

【００５３】

コーパス入力手段１１は、コーパス２を入力する処理手段である。入力されるコーパス２は、テキストデータであって、例えば電子化された大量の新聞記事データ、論文データなどである。

【００５４】

タグ登録手段１２は、ユーザが入力装置２２を介して、所定の言語解析処理の結果となる言語情報の個々の分類先に対応する分類タグを指定すると、指定された分類先および分類タグを入力してタグ記憶手段１３に格納する処理手段である。

【００５５】

分類タグは、ＳＧＭＬ（Standard Generalized Markup Language）形式にもとづいて例えば＜PERSON＞＜/PERSON ＞、＜LOCATION＞＜/LOCATION ＞のように記述される属性情報である。一対の分類タグに囲まれた要素（文字列）が、その分類タグに対応する言語情報

（分類先）が付与される対象となることを示す。
【００５６】
タグ付与手段１４は、コーパス２のテキストデータを表示装置２１に表示し、表示装置２１に表示されたテキストデータ上において、ユーザによって選択された文字列の前後に分類タグを挿入し、分類タグが挿入されたテキストデータの分類タグが挿入された箇所を含む所定の範囲の前後に、範囲指定タグを挿入する処理手段である。
【００５７】
範囲指定タグは、教師データとして使用する範囲を示すタグであって、分類タグと同様にＳＧＭＬ形式で記述される属性情報であり、例えば、＜UC＞＜/UC ＞のように記述される。
【００５８】
タグ付与手段１４は、分類タグが挿入されたテキストデータにおいて、ユーザによって指定された前記分類タグが挿入された箇所を含む範囲の前後に前記範囲指定タグを挿入し、または、分類タグが挿入された箇所を含む所定の範囲を所定の範囲指定規則にもとづいて指定し、前記指定された範囲の前後に前記範囲指定タグを挿入する。
【００５９】
所定の範囲指定規則として、例えば、ユーザが分類タグを付与した箇所を含む一または複数の文もしくは段落や、ユーザが分類タグを付与した箇所と同一の文字列を含む文、ユーザが分類タグを付与した箇所の前方または後方に連なる所定の単語数もしくは文字数の範囲などをユーザが指定した範囲とみなすような規則を予め備えておく。また、分類タグを含む同一文については句点で文の認識を行い、分類タグを含む同一段落内については、改行、字下げ、空行などで認識を行い、または同一行、前後に所定の行数の行を含む部分などとする規則を設けておく。
【００６０】
ユーザ範囲抽出手段１６は、分類タグおよび範囲指定タグが挿入されたテキストデータから、範囲指定タグに囲まれたデータを、教師データを生成するためのユーザ範囲データとして抽出する処理手段である。
【００６１】
教師データ変換手段１７は、ユーザ範囲データを所定の単位（形態素、文字、文字列、単語、文節、文など）ごとに切り出し、ユーザ範囲データから分類タグに囲まれた文字列を検出し、切り出し部分のうち前記検出した文字列に対応する部分に前記分類タグに対応する分類先を前記単位ごとに付与し、各単位のデータを、解を前記分類先とする教師データに変換する処理手段である。
【００６２】
教師データ変換手段１７は、分類タグが付与されて検出された文字列が、複数の教師データの切り出し単位からなる場合に、分類先かつ文字列におけるその単位の位置を示す情報を付加したものを、単位ごとに付与する。
【００６３】
例えば、分類タグが付与された文字列が単語であり、教師データとして切り出される単位が文字である場合に、教師データ変換手段１７は、文字列の先頭の文字には、その分類先と文字列の先頭であることを示す分類先「B-…」、それ以外の文字には、その分類先と文字列の先頭以外の文字であることを示す分類先「I-…」を付与する。
【００６４】
素性抽出手段１１０は、教師データから所定の種類の素性を抽出し、教師データの切り出し単位について、素性の集合と付与された分類先との組を生成する処理手段である。
【００６５】
素性抽出手段１１０は、教師データに対して形態素解析を行い素性を抽出し、または、教師データから所定の文字または文字列を切り出して素性とする。
【００６６】
規則登録手段１８は、ユーザによって定義されたユーザ範囲指定規則を規則記憶手段１

９に記憶する処理手段である。

【００６７】

表示装置２１は、ユーザが登録された分類タグを選択できる選択項目、コーパス入力手段１１により入力されたコーパス（テキストデータ）２を表示して、分類タグや範囲指定タグを付加する箇所を指定できる指定項目などを備えるタグ付与画面を表示する装置である。

【００６８】

入力装置２２は、種々のデータやユーザ指示などを入力する装置であって、タグ付与画面に表示されたテキストデータ上で範囲や位置などを指定し、選択項目を指定するものである。例えば、マウス、カーソルキーおよび実行キーを備えるキーボードなどである。

【００６９】

言語解析処理装置４は、教師データ作成装置１により作成された教師データを入力して機械学習法を用いた所定の言語解析処理を行う装置である。言語解析処理装置４は、機械学習手段４２、学習結果記憶手段４３、データ入力手段４４、素性抽出手段４５、解推定手段４６、タグ付与手段４７、解析結果表示処理手段４８、および表示装置４９を備える。

【００７０】

機械学習手段４２は、教師データの前記素性の集合と前記分類先との組を利用して、各単位について、素性の集合の場合にどのような分類先になりやすいかを学習し、学習の結果を学習結果記憶手段４３に記憶しておく処理手段である。

【００７１】

データ入力手段４４は、言語解析処理の対象とするテキストデータを入力する処理手段である。

【００７２】

素性抽出手段４５は、前記入力データから所定の解析処理または切り出し処理により、所定の単位（形態素、文字、単語、文節など）について所定の種類の素性を抽出する手段である。

【００７３】

解推定手段４６は、学習結果記憶手段４３に記憶された前記学習結果を利用して、入力データの所定の単位のデータについてその素性の場合になりやすい分類先を推定する処理手段である。

【００７４】

タグ付与手段４７は、推定された分類先に対応する分類タグを、入力データの推定の対象となった単位に対応する文字列の前後に挿入する処理手段である。

【００７５】

解析結果表示処理手段４８は、、分類タグごとに色または表示態様を違えて表示するように定めた所定の表示規則をもとに、入力データの分類タグが挿入された箇所と分類タグが挿入されていない箇所とを違えて表示装置４９に表示する処理手段である。

【００７６】

なお、言語解析処理装置４は、機械学習法として、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の手法を用いる。

【００７７】

言語解析処理装置４が、サポートベクトルマシン法を用いる場合には、機械学習手段４２では、教師データから解となりうる分類先を特定し、その分類先を正例と負例に分割し、所定のカーネル関数を用いたサポートベクトルマシン法を実行する関数にしたがって素性の集合を次元とする空間上で正例と負例の間隔を最大にして正例と負例を超平面で分割する超平面を求め、その超平面を学習結果とし、その超平面を学習結果記憶手段４３に記憶する。そして、解推定手段４６では、学習結果記憶手段４３に記憶されている学習結果の超平面を利用して、入力データの素性の集合がこの超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果に基づいて定まる分類先を

、入力データの素性の集合の場合になりやすい分類先と推定する。

〔第１の実施例〕

第１の実施例として、言語解析処理装置４で機械学習法を用いて固有表現抽出処理を行う場合に、教師データ作成装置１で言語解析処理装置４が使用する教師データを作成する処理を説明する。

【００７８】

固有表現抽出処理とは、テキストデータから地名、人名、組織名、数値表現などの固有な表現を抽出する処理をいう。固有表現抽出処理において解析結果となる分類先は、例えば地名、人名、組織名、日付表現、時間表現、金額表現、割合表現などである。教師データには、これらの分類先それぞれに対応する分類ラベルが付与される。

【００７９】

図２は、教師データ作成処理の処理フローを示す図である。

【００８０】

教師データ作成装置１のタグ登録手段１２は、ユーザが、入力装置２２を介して、以下のような固有表現抽出処理の分類先とそれに対応する分類タグを指定すると、ユーザが指定した分類先およびその分類タグ（開始タグと終了タグ）を入力してタグ記憶手段１３に記憶する（ステップＳ１０）。

【００８１】

＜PERSON＞＜/PERSON ＞：分類先＝人名、
＜LOCATION＞＜/LOCATION ＞：分類先＝地名、
＜ORGANIZATION＞＜/ORGANIZATION ＞：分類先＝組織名、
＜ARTIFACT＞＜/ARTIFACT ＞：分類先＝固有物名、
＜DATE＞＜/DATE ＞：分類先＝日付表現、
＜TIME＞＜/TIME ＞：分類先＝時間表現、
＜MONEY ＞＜/MONEY＞：分類先＝金額表現、
＜PERCENT ＞＜/PERCENT＞：分類先＝割合表現、…。

【００８２】

本例では、付与する分類ラベルを文字単位に付与した教師データを作成する。例えば、＜PERSON＞＜/PERSON ＞分類タグが対応する分類先「人名」の分類ラベルは、先頭文字を示す「B-」または先頭以外の文字を示す「I-」を付けて、「B-PERSON」、「I-PERSON」とする。また、分類先に該当しない文字に付与するラベルとして、「OTHER 」を登録する。

【００８３】

また、固有表現抽出処理の分類先として字種を用いる場合には、以下のような分類先および分類タグをタグ記憶手段１３に格納する。

【００８４】

＜KANJI ＞＜/KANJI＞：分類先＝漢字、
＜KATAKANA＞＜/KATAKANA ＞：分類先＝カタカナ、
＜ALPHABETIC＞＜/ALPHABETIC ＞：分類先＝英字、
＜NUMERIC ＞＜/NUMERIC＞：分類先＝数字。

【００８５】

そして、コーパス入力手段１１が、固有表現抽出処理の分類先が付与されていないテキストデータで構成されるコーパス２を入力すると（ステップＳ１１）、タグ付与手段１４は、コーパス２のテキストデータを表示しユーザにタグ付与操作を促すタグ付与画面を表示装置２１に表示する（ステップＳ１２）。

【００８６】

図３は、タグ付与画面の例を示す図である。タグ付与画面１００は、コーパス２のテキストデータを表示して分類タグを付加する箇所を指定できる指定項目１０１、タグ記憶手段１３に格納された分類先の一覧表示から任意の分類先を選択できる選択項目１０３などで構成される。

【００８７】

ユーザによって、分類先を付与したい箇所および付与する分類先が指定されたら（ステップＳ１３）、タグ付与手段１４は、タグ付与画面１００で指定された箇所に対応する文字列の前後に選択された分類タグを挿入する（ステップＳ１４）。

【００８８】

例えば、入力されたコーパス２に、テキストデータ「…日本の首相は小泉さんです。小泉さんはいつも思いきったことをしています。…」が含まれていたとする。図３（Ａ）に示すように、ユーザが、タグ付与画面１００の指定項目１０１に表示されたテキストデータ上で、マウスドラッグ操作などにより、分類先を付与する単語「日本」を指定する。さらにマウスの右ボタンクリック操作を行って表示させた選択項目１０３から、マウス左ボタンクリック操作などにより分類先「地名」を選択する。同様に、指定項目１０１で単語「小泉」を指定し、選択項目１０３から分類先「人名」を選択する。

【００８９】

タグ付与手段１４は、タグ付与画面１００で指定された箇所に対応するテキストデータ中の文字列の前後に、選択された分類タグを挿入する。分類タグが付与されたテキストデータは以下のようになる。

「…＜LOCATION＞日本＜/LOCATION ＞の首相は＜PERSON＞小泉＜/PERSON ＞さんです。小泉さんはいつも思いきったことをしています。…」

さらに、ユーザによって、指定項目１０１で分類先を付与する作業を行い教師データとして使用する範囲が指定されると（ステップＳ１５）、タグ付与手段１４は、タグ付与画面１００で指定された範囲に対応するテキストデータの文字列の前後に範囲指定タグの開始タグおよび終了タグを付加する（ステップＳ１６）。例えば、図３（Ｂ）に示すように、ユーザが、マウスドラッグにより文「日本の首相は小泉さんです。」を範囲として指定したとする。タグ付与手段１４は、指定された範囲に対応するテキストデータの文字列の前後に範囲指定タグを挿入する。範囲指定タグが付与されたテキストデータは以下のようになる。

「…＜UC＞＜LOCATION＞日本＜/LOCATION ＞の首相は＜PERSON＞小泉＜/PERSON ＞さんです。＜/UC ＞小泉さんはいつも思いきったことをしています。…」

一方、ユーザが、分類先を付与した後、教師データとして使用する範囲を指定しなかった場合には、タグ付与手段１４は、指定項目１０１で分類先が付与された箇所を含む所定の箇所をユーザが選択した範囲とみなし、その範囲の前後に範囲指定タグを付加する（ステップＳ１７）。例えば、タグ付与手段１４は、テキストデータ中の分類タグが付与された文字列に単語の前後に連なる所定の文字数や単語数などの範囲を、ユーザが選択した範囲とみなし、みなした範囲の前後に範囲指定タグを付加する。

【００９０】

そして、タグ付与手段１４は、テキストデータに分類タグおよび範囲指定タグを付加したテキストデータ（タグ付きコーパス）をコーパス記憶手段１５に記憶する（ステップＳ１８）。

【００９１】

その後、ユーザ範囲抽出手段１６は、コーパス記憶手段１５のタグ付きコーパスから、範囲指定タグの開始タグ＜UC＞と終了タグ＜/UC ＞とに囲まれた範囲のテキストデータ（ユーザ範囲データ）を抽出する（ステップＳ１９）。

【００９２】

そして、教師データ変換手段１７は、抽出されたテキストデータを所定の単位（ここでは文字単位とする）に分割し、抽出されたテキストデータから分類タグに囲まれた文字列を検出し、各単位（文字）のうち分類タグが付与されている文字に分類タグに対応する分類ラベルを付与し、分類タグが付与されていない文字に分類先がないことを示す分類ラベルを付与して、教師データとする（ステップＳ１１０）。

【００９３】

図４は、教師データの例を示す図である。例えば、教師データとして、範囲指定タグに囲まれたテキストデータ「＜UC＞＜LOCATION＞日本＜/LOCATION ＞の首相は＜PERSON＞小

泉＜/PERSON ＞さんです。＜/UC ＞」が抽出されたとする。教師データ変換手段１７は、例えば、テキストデータの分類タグ＜PERSON＞と＜/PERSON ＞に囲まれた文字列「小、泉」の先頭文字「小」に、分類先「人名」の先頭を示す分類ラベル「B-PERSON」を、同じく次の文字「泉」に分類先「人名」の先頭以外を示す分類ラベル「I-PERSON」を付与する。また、テキストデータのうち分類タグに囲まれていない部分「の、首、相、は、さ、ん、で、す、。」について、各文字にユーザが指定した分類先に該当しない旨を示す分類ラベル「O 」を付与する。

【００９４】

そして、素性抽出手段１１０により、教師データに対して形態素解析処理を行い、所定の単位（例えば文字）ごとの素性を抽出し、素性の集合と分類ラベルとの組を生成する（ステップＳ１１１）。

【００９５】

形態素解析処理は、例えば、以下の参照文献１に示す形態素解析システム「茶筌（ChaSen）」を用いる。形態素解析システム「茶筌（ChaSen）」は、コスト幅のオプションを設定することにより冗長な解析出力が可能な形態素解析であって、コストとしてマルコフモデルにもとづいてテキストデータから推定された対数尤度を用いるコスト最小法を用いたものである。例えば、文「学校へ行く」を入力すると、以下のように、各行に一個の単語が入るように分割され、各単語に読みや品詞などの言語情報が付与された出力結果を得ることができる。［参照文献１：http://chasen.aist-nara.ac.jp/index.html.ja ］

「学校：ガッコウ，　学校，名詞－一般；
へ：　へ，へ，　助詞－格助詞－一般；
行く：イク，行く，　動詞－自立　五段・カ行促音便　基本形；
EOS」

なお、素性抽出手段１１０として、既知の他の形態素解析処理装置を用いてもよい。

【００９６】

また、教師データ作成処理として、ステップＳ１６の処理もしくはステップＳ１７の処理のいずれか一方の処理のみを行ってもよく、また、ステップＳ１６およびステップＳ１７の処理の両方を行ってもよい。

【００９７】

図５は、教師データの各文字の素性と付与された分類ラベル（解）との組の例を示す図である。素性として、例えば、品詞情報（名詞、固有名詞、人名、姓、などの分類）、形態素における文字の位置情報（先頭、それ以外などの分類）、字種情報（漢字、カタカナ、英字、数字などの分類）、分類先などが抽出される。

【００９８】

図６は、言語解析処理の処理フローを示す図である。

【００９９】

言語解析処理装置４は、機械学習手段４２では、素性の集合と分類ラベルの組を利用して、各単位（文字）について、その素性の集合の場合にどのような分類先になりやすいかを学習し（ステップＳ２０）、学習結果を学習結果記憶手段４３に記憶する（ステップＳ２１）。

【０１００】

機械学習手段４２は、例えば、図５に示す各文字の素性と分類ラベルとの組において、文字「小」についての学習には、破線で示す矩形で囲まれた素性の集合を用いて行う。

【０１０１】

ここで、機械学習法としては、多分類に対応できる拡張したサポートベクトルマシン法を用いる。

【０１０２】

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。このとき、２つの分類が正例と負例からなるものとすると、学習データにおける正例と負例の間隔（マージン）が大きいものほど、オープンデータ

で誤った分類をする可能性が低いと考えられ、このマージンを最大にする超平面を求め、求めた超平面を用いて分類を行う。

【０１０３】

図７は、サポートベクトルマシン法の最大マージンを説明するための図である。図７に示すように、ある空間で求める分離超平面（実線で示す）と、分類超平面に平行かつ等距離にある超平面（破線で示す）の距離（マージン）が最大になるような分離超平面を求める。

【０１０４】

サポートベクトルマシン法では、通常、学習データにおいて、マージンの内部領域に小量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形にする拡張（カーネル関数の導入）がなされたものが用いられる。このような拡張された方法は、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、２つの分類を判別することができる。

【０１０５】

【数１】

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

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

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

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_i （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　（ｘ≧０）

－１（otherwise）　式（２）

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

【０１０６】

【数２】

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

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

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

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、本例では以下の多項式（６）を用いる。Ｃ、ｄは、実験的に設定される定数である。

【０１０７】

$K(x, y) = (x \cdot y + 1)^d$　式（６）

なお、サポートベクトルマシンは、正例・負例の二値分類であるため、ワン・バーサス・レスト（One v.s. Rest ）法、ペア・ワイズ(Pair Wise)法などの手法を用いて二値分類を多値分類に拡張する。

【０１０８】
ワン・バーサス・レスト（One v.s. Rest ）法では、例えば３つの分類先ａ、ｂ、ｃがある場合に、「ａとその他」、「ｂとその他」、「ｃとその他」という３つの組の二値分類器（ある分類先か、それ以外の分類先か）を用意し、それぞれをサポートベクトルマシンで学習する。そして、解である分類先を推定する場合には、３つのサポートベクトルマシンの学習結果を利用する。推定するべき入力データが、これらの３つのサポートベクトルマシンでは、どのように推定されるかをみて、３つのサポートベクトルマシンのうち、その他でない側（正例）に分類されかつサポートベクトルマシンの分離平面から最も離れた場合のものの分類先を、求める解とする。
【０１０９】
ペア・ワイズ(Pair Wise)法では、ｋ個の分類先から任意の２つの分類先についての二値分類器を${}_kC_2$個用意して、分類先同士の総当たり戦を行い、このうち最も分類先として選ばれた回数が多い分類先を求める解とする。
【０１１０】
機械学習の学習終了後、データ入力手段４４では、言語解析の対象のテキストデータを入力する（ステップＳ２２）。素性抽出手段４５では、教師データ作成処理のステップＳ１１１の処理と同様に、入力されたテキストデータ（入力データ）に対して形態素解析を行い、所定の単位（例えば文字）ごとの素性を抽出する（ステップＳ２３）。
【０１１１】
そして、解推定手段４６では、学習結果記憶手段４３に記憶された学習結果を利用して、入力データの所定の単位（文字）について、その素性の場合になりやすい分類ラベルを推定する（ステップＳ２４）。
【０１１２】
そして、タグ付与手段４７は、解と推定された分類ラベルに対応する分類タグを、入力データの該当する文字または文字列の前後に挿入する（ステップＳ２５）。図８（Ａ）は、入力データの文字ごとに推定された解（分類ラベル）の例を示す図、図８（Ｂ）は、分類タグが付与された入力データの例を示す図である。
【０１１３】
解析結果表示処理手段４８では、分類タグが付加された入力データを、所定の表示規則に従った表示態様で表示装置４９に表示する（ステップＳ２６）。ここで、分類タグ＜PERSON＞＜/PERSON ＞で囲まれた文字列を青色で表示し、＜LOCATION＞＜/LOCATION ＞で囲まれた文字列を赤色で表示する表示規則がある場合に、解析結果表示処理手段４８は、「森」を青色にして「森さんが前の首相です。」を表示装置４９に表示する。これにより、特定の固有表現の抽出結果を分かりやすく表示することができる。
【０１１４】
別の例として、固有表現抽出処理の分類先として、ユーザが、「賛成語」と「反対語」を指定した場合の処理を説明する。
【０１１５】
例えば、新聞記事のうち社説などの論調は、新聞社によって異なることが多い。ユーザが、このような論調を新聞社や社説ごとに整理したいと考える場合に、例えば賛成や反対を示す表現が重要となるため、新聞記事データなどのコーパスから、以下のような賛成語と反対語のさまざまな表現を抽出できれば便利である。
「賛成語」：支持した、賛成した、同意した、了承した、…、
「反対語」：反対した、同調しなかった、…。
【０１１６】
この「賛成語」と「反対語」のように、ユーザが定義した分類先にもとづいて所定の表現を抽出する場合に、タグ登録手段１２は、ユーザが指定した以下のような分類先と分類タグとをタグ記憶手段１３に登録する。
＜APPROVAL＞＜/APPROVAL ＞：分類先＝賛成語
＜DISAPPROVAL ＞＜/DISAPPROVAL＞：分類先＝反対語

なお、各分類ラベルに単位内での先頭文字を示す「B-」または先頭以外の文字を示す「I-」の区別を付け、分類先に該当しない旨の分類ラベルとして「OTHER 」を登録する。

【０１１７】

そして、タグ付与手段１４は、コーパス入力手段１１が入力したコーパス２のテキストデータ「…日本は米国を支持したが、フランスは反対した。ドイツも反対した。…」を含むタグ付与画面を表示する。

【０１１８】

図９は、タグ付与画面の例を示す図である。ユーザが、タグ付与画面１００の指定項目１０１に表示されたテキストデータの「反対した」を選択し、選択項目１０３から分類先「反対」を選択すると、タグ付与手段１４は、テキストデータの文字列「反対した」の前後に分類タグ＜DISAPPROVAL ＞＜/DISAPPROVAL＞を挿入する。

【０１１９】

その後ユーザがユーザ範囲を指定しなかった場合には、分類タグが付与された文「日本は米国を支持したが、フランスは反対した。」をユーザが指定した範囲とみなし、この文の前後に範囲指定タグ＜UC＞＜/UC ＞を挿入する。

【０１２０】

タグ付与手段１４は、コーパス記憶手段１５に、以下のタグが付与されたテキストデータを含むタグ付きコーパスを記憶する。

「＜UC＞日本は米国を＜APPROVAL＞支持した＜/APPROVAL ＞が、フランスは＜DISAPPROVAL ＞反対した＜/DISAPPROVAL＞。＜/UC ＞ドイツも反対した。」

ユーザ範囲抽出手段１６が、タグ付きコーパスから範囲指定タグに囲まれた部分「＜UC＞日本は米国を＜APPROVAL＞支持した＜/APPROVAL ＞が、フランスは＜DISAPPROVAL ＞反対した＜/DISAPPROVAL＞。」を抽出すると、教師データ変換手段１７は、抽出されたテキストデータの各文字に分類タグに対応する分類ラベルを付与して教師データとする。図１０は、教師データの例を示す図である。

【０１２１】

言語解析処理装置４の各処理手段の処理は、既に説明した処理と同様である。データ入力手段４４が入力データ「ロシアは反対した。」を入力した場合に、素性抽出手段４５は入力データの文字ごとに素性を抽出し、機械学習手段４２は、各文字ごとの解（分類ラベル）を推定する。タグ付与手段４７は、図１１に示すように、入力データに分類ラベルが付与された文字列の前後に、その分類ラベルに対応する分類タグを挿入する。

【０１２２】

解析結果表示処理手段４８は、分類タグ＜APPROVAL＞＜/APPROVAL ＞で囲まれた文字列を青色で、分類タグ＜DISAPPROVAL ＞＜/DISAPPROVAL＞で囲まれた文字列を赤色で表示するという表示規則を備えている場合に、入力データ中の分類先「反対」の分類タグで囲まれた「反対した」を赤色で表示する。

〔第２の実施例〕

第２の実施例として、言語解析処理装置４で機械学習法を用いて照応解析処理を行う場合に、教師データ作成装置１で言語解析処理装置４が使用する教師データを作成する処理を説明する。

【０１２３】

照応解析処理とは、テキストデータの文の代名詞、定名詞、指示詞などの語（指示表現という）が、文の並びである文脈中の別の語（指示先という）と同じ対象を指示するという現象を解析する処理である。解析結果として、指示先と指示表現との関係を同定する必要がある。例えば、解析結果となる言語情報の分類先としては、以下のような分類先が必要となる。

１）指示先となる対象（個体）の最初の出現（「固体導入」）であるか、否か、
２）前方の一番目に近い名詞句が指示先（「名詞（１番目）を指示」）であるか、否か、
３）前方の二番目に近い名詞句が指示先（「名詞（２番目）を指示」）であるか、否か。

【０１２４】

これらの分類先に対応する照応タグ（開始タグと終了タグ）として＜ref ｎ＞＜/ref＞のように記述するタグを用意しておき、ユーザが選択した同一対象を指示する照応関係の単語に付与する。＜ref ｎ＞のｎには、同一対象に同一値が設定される。

【０１２５】

教師データ作成装置１のコーパス入力手段１１は、テキストデータで構成されるコーパス２を入力する。例えば、入力されたコーパス２に、テキストデータとして以下のデータが含まれていたとする。

「おじいさんがすんでいました。おじいさんは山にいきました。そこには大きな木がたっていました。木には小鳥の巣がありました。…」。

【０１２６】

ユーザが、タグ付与画面のテキストデータ上でマウスドラッグ操作により、第１文の単語「おじいさん」を指示先として選択し、第２文の単語「おじいさん」を指示表現として選択し照応タグ＜ref 0 ＞を付与する。タグ付与手段１４は、画面上で指定された文字列の前後に照応タグの開始タグ＜ref 0 ＞および終了タグ＜/ref＞を挿入する。

【０１２７】

同様に、ユーザが第２文の単語「山」を指示先として選択し、第３文の単語「そこ」を指示表現として選択して照応タグ＜ref 1 ＞を選択すると、タグ付与手段１４は、それぞれの文字列の前後に＜ref 1 ＞＜/ref＞を挿入する。

【０１２８】

照応タグが付与されたテキストデータは、以下のようになる。

「＜ref 0 ＞おじいさん＜/ref＞がすんでいました。＜ref 0 ＞おじいさん＜/ref＞は＜ref 1 ＞山＜/ref＞にいきました。＜ref 1 ＞そこ＜/ref＞には大きな木がたっていました。木には小鳥の巣がありました。…」。

【０１２９】

ここで、＜ref 0 ＞が付与された二つの「おじいさん」、および＜ref 1 ＞が付与された「山」および「そこ」が、それぞれで同一の対象であることを示す。

【０１３０】

その後、タグ付与画面で、ユーザは第１文から第３文までしか照応関係をチェックしなかったとする。タグ付与手段１４は、テキストデータ中の照応タグが付与された文を含む範囲を、ユーザがタグ付与作業を行った範囲とみなして、前後に範囲指定タグの開始タグ＜UC＞および終了タグ＜/UC ＞を挿入する。

「＜UC＞＜ref 0 ＞おじいさん＜/ref＞がすんでいました。＜ref 0 ＞おじいさん＜/ref＞は＜ref 1 ＞山＜/ref＞にいきました。＜ref 1 ＞そこ＜/ref＞には大きな木がたっていました。＜/UC ＞木には小鳥の巣がありました。…」。

【０１３１】

そして、タグ付与手段１４は、照応タグが付与されたテキストデータ（タグ付きコーパス）をコーパス記憶手段１５に記憶する。

【０１３２】

その後、ユーザ範囲抽出手段１６は、コーパス記憶手段１５のタグ付きコーパスから、範囲指定タグの開始タグ＜UC＞と終了タグ＜/UC ＞とに囲まれた範囲のテキストデータを抽出する。

「＜UC＞＜ref 0 ＞おじいさん＜/ref＞がすんでいました。＜ref 0 ＞おじいさん＜/ref＞は＜ref 1 ＞山＜/ref＞にいきました。＜ref 1 ＞そこ＜/ref＞には大きな木がたっていました。＜/UC ＞」。

【０１３３】

教師データ変換手段１７は、抽出されたテキストデータを所定の単位（単語）に分割し、テキストデータの照応タグが付与された単語を検出し、検出した単語に分類ラベルを付与する。例えば、抽出されたテキストデータで照応タグ＜ref 0 ＞が付与された単語を検出し、最初に出現した単語（第１文の「おじいさん」）に分類ラベル「個体導入」を付与し、次の単語（第２文の「おじいさん」）に分類ラベル「おじいさん（１番目）を指示」

を付与する。

【０１３４】

同様に、照応タグ＜ref １ ＞が付与された単語を検出し、最初に出現した単語（第２文の「山」）に分類ラベル「個体導入」を付与し、次の単語（第３文の「そこ」）に分類ラベル「山（１番目）を指示」を付与する。

【０１３５】

なお、さらに照応タグ＜ref １ ＞が付与された単語を検出した場合には、その単語（例えば「そこ」）に分類ラベル「山（２番目）を指示」を付与する。

【０１３６】

そして、教師データ変換手段１７は、抽出されたテキストデータの各単位を教師データとする。図１２は、教師データの例を示す図である。

【０１３７】

さらに、素性抽出手段１１０は、教師データに対して形態素解析、構文解析などの処理を行い、品詞情報の他、照応解析に関する所定の種類の素性を抽出する。

【０１３８】

形態素解析処理は、例えば、参照文献１に示す形態素解析システム「茶筌（ChaSen）」を用いて行い、品詞情報などの素性を抽出する。また、構文解析処理は、例えば、参照文献２に示す言語解析システム「南瓜（CaboCha ）を用いて行い、文節または文節間の係り受けの情報などの素性を抽出する。［参照文献２：ＳＶＭに基く日本語係り受け解析器 CaboCha「南瓜」、http://cl.aist-nara.ac.jp/` taku-ku/software/cabocha/ ］

また、抽出する素性の種類は、以下のとおりである。

素性(1) ：指示表現、
素性(2) ：指示先の表現、もしくは、個体導入か、
素性(3) ：指示表現と指示先との距離、何文節離れているか
（個体導入の場合「０文節離れている」とする）
素性(4) ：指示表現と指示先の意味的整合性があっているかどうか、もしくは、
個体導入か、
素性(5) ：指示表現が係る動詞がその指示表現のある格にとりうる意味と指示先の意味的整合性があっているかどうか、もしくは、個体導入か、
素性(6) ：前方に同一名詞があるか、否か

図１３および図１４に、教師データの各単語の素性(1) ～素性(6) として抽出された素性を示す。図１３は、教師データのうち＜ref ０ ＞および＜ref １ ＞に関するデータについて、前出の「１）指示先となる対象（個体という）の最初の出現である固体導入であるか、否か」という分類先の学習を行う場合の素性と分類ラベルの組の例を示す。図１４は、同じデータについて前出の「２）前方の一番目に近い名詞句が指示先であるか、否か」という分類先の学習を行う場合の素性と分類ラベルの組の例である。

【０１３９】

ここで、素性(4) の意味的整合性は、あらかじめ人手によって作成しておいた規則にもとづいて判断する。例えば、以下のような規則を作成しておく。

「1)指示表現と指示先の表現が完全に一致する場合、または、
2)指示先の表現が指示表現を含む場合、または、
3)指示先の表現が指示表現の下位語の場合、または、
4)指示表現に対して予め作成した表現の語のリストの中に指示先の表現がある場合に、
→ 指示表現と指示先の意味的整合性がある」

下位語とは、他の概念に包括される関係にある概念の語をいう。「鳥」の下位語として、例えば「にわとり」、「からす」、「つる」などが該当する。

【０１４０】

また、ある指示表現に対して対応しうる表現の語のリストを作成しておく。例えば、指示表現となる指示詞「そこ」は場所を意味する語を指示しうるので、山、畑、海岸、公園などの場所を意味する語のリストを作成しておく。

【０１４１】
図１３に示す二つの「おじいさん」の意味的整合性は、指示表現と指示先の表現が完全に一致する場合に相当し、意味的整合性があると判断する。また、図１４に示す「山」と「そこ」については、「山」と「そこ」の意味的整合性は、指示表現「そこ」のリストに「山」が含まれているので、「指示表現に対して予め作成した表現の語のリストの中に指示先の表現がある場合」に相当し、意味的整合性はあると判断する。
【０１４２】
また、素性(5) の「意味的整合性」は、動詞の格フレーム辞書および名詞意味辞書を用意し、指示表現がかかる動詞がとりうる意味と指示先の表現の語の意味とを利用して判断する。
【０１４３】
動詞の格フレーム辞書は、以下に示すように、その動詞が、どのような格を持ち、その格がどのような意味の表現をとりうるかを記述するデータである。
「いく：がー動物、にー場所；　たつ：がーもの、にー場所、…」。
【０１４４】
この例では、「いく」はガ格とニ格を持ち、ガ格では動物を意味する表現を、ニ格では場所を意味する表現をとりうることを、また、「たつ」はガ格とニ格を持ち、ガ格ではものを意味する表現を、ニ格では場所を意味する表現をとりうることを示している。
【０１４５】
名詞意味辞書は、以下のように、名詞ごとに、その名詞がどういう意味になりうるかを記述したデータである。
「おじいさん：人、動物、もの；　山：場所、もの、…。」
この例では、おじいさんは、人、動物、ものを意味し、山は、場所、ものをそれぞれ意味することを示す。
【０１４６】
さらに、素性(5) の「指示表現が係る動詞がその指示表現のある格にとりうる意味と指示先の意味的整合性があっているかどうか」は、動詞の格フレーム辞書を用いて、指示表現が係る動詞について指示表現のある格にとりうる意味を把握し、指示先の表現の意味を名詞意味辞書を用いて把握する。そして、それらの意味が一致する場合があるかどうかを調べ、一致する場合があれば、意味的整合性があっていると判断する。
【０１４７】
例えば、図１３の二つの「おじいさん」について、構文解析の結果、２番目の「おじいさん」が係る動詞は「いく」であり、この「おじいさん」はガ格でことがわかる。そこで、動詞の格フレーム辞書から、指示表現が係る動詞（いく）がその指示表現のある格（ガ格）にとりうる意味は「動物」であることがわかる。また、指示先の表現の１番目の「おじいさん」は、名詞意味辞書から、その意味「人、動物、または、もの」であることがわかる。そして、指示先の表現の１番目の「おじいさん」が「動物」の意味である場合に、指示表現が係る動詞がその指示表現のある格にとりうる意味と指示先の意味とが一致するので、意味的整合性があっていると判断する。
【０１４８】
また、図１４に示す「山」と「そこ」ついて、構文解析の結果、「そこ」が係る動詞は「たつ」であり、２番目の「おじいさん」はニ格であることがわかる。そこで、動詞の格フレーム辞書から、指示表現が係る動詞（たつ）がその指示表現のある格（ニ格）にとりうる意味は「場所」であることがわかり、指示先の表現の「山」は、名詞意味辞書から、その意味が「場所、またはもの」であることがわかる。そして、指示先の表現の「山」が「場所」の意味である場合に、指示表現が係る動詞がその指示表現のある格にとりうる意味と指示先の意味とが一致するので、意味的整合性があっていると判断する。
【０１４９】
機械学習手段４２は、これらの各単位（名詞句）について、その名詞句の素性と分類ラベルとの組を利用して、各名詞句について、その素性の集合の場合にどのような分類先に

なりやすいかを学習し、その学習結果を学習結果記憶手段４３に記憶する。
【０１５０】
機械学習手段４２での学習終了後、データ入力手段４４は、言語解析処理の対象としたいテキストデータを入力する。素性抽出手段４５は、素性抽出手段１１０と同様に、入力された文章データ（入力データ）の形態素解析および構文解析を行い、名詞句について素性を抽出する。
【０１５１】
そして、解推定手段４６は、学習結果記憶手段４３に記憶しておいた学習結果を参照し、入力データの各単語について、その素性の場合に最も分類されやすい分類ラベルを推定する。分類ラベルは、
１）「固体導入」、それ以外か、
２）「前方の１番目に近い名詞句が指示先（名詞（１番目）を指示）」、それ以外か、
３）「前方の２番目に近い名詞句が指示先（名詞（２番目）を指示）」、それ以外か、
のそれぞれについて二値分類を推定し、その結果をもとに指示先を推定する。
【０１５２】
入力データが、「おばあさんは畑へいきました。畑には大根がいっぱいうわっていました。」であるとする。
【０１５３】
形態素解析および構文解析の結果抽出した入力データの名詞句の素性(1) ～素性(6) を用いて、「１）固体導入か、それ以外か」の分類ラベルを推定する場合に、使用する素性は以下のようになる。
「おばあさん：おばあさん、固体導入、０個、固体導入、固体導入、なし；
畑（１番目）：畑、固体導入、０個、固体導入、固体導入、なし；
畑（２番目）：畑、固体導入、０個、固体導入、固体導入、あり；
大根：大根、固体導入、０個、固体導入、固体導入、なし；」
機械学習手段４２は、おばあさん、畑（１番目）、大根の分類ラベルを「固体導入」と推定し、畑（２番目）の分類ラベルを「それ以外」と推定する。
【０１５４】
また、２番目の「畑」について、「２）前方の１番目に近い名詞句が指示先か、それ以外か、」の分類ラベルを推定する場合に、使用する素性は、以下のようになる。
「畑（２番目）：畑、畑、２個、整合性有り、整合性有り、同一名詞あり」
機械学習手段４２は、「畑（２番目）」の分類ラベルは「前方の１番目に近い名詞句が指示先」であると推定し、最終的に、各単語の分類ラベルを以下のように推定する。
「おばあさん（１番目）：個体導入、
畑（１番目）：個体導入、
畑（２番目）：畑（１番目）を指示、
大根（１番目）：個体導入」
そして、機械学習手段４２は、各単語の推定した分類ラベルから、「畑（１番目）」と「畑（２番目）」とが照応関係であると解析する。
【０１５５】
その後、タグ付与手段４７は、照応関係を持つと推定した入力データの単語（畑）の前後に、同じ数字の照応タグを挿入する。
「おばあさんは＜ref 0 ＞畑＜/ref＞へいきました。＜ref 0 ＞畑＜/ref＞には大根がいっぱいうわっていました。」
その後、解析結果表示処理手段４８は、所定の表示規則に従って、同じ数字の照応タグに囲まれた名詞を同じ色で表示するなどして入力データを表示装置４９に表示する。これにより、同一の指示対象についての照応関係を分かりやすく表示することができる。
〔第３の実施例〕
第３の実施例として、言語解析処理の一つである要約処理について、機械学習法を用いてユーザの指向に適応する要約処理を説明する。要約処理とは、文章データを、その内容

を表わすために重要と考えられる文（重要文という）を用いて要約する処理をいう。

【０１５６】

図１５は、要約処理装置６の構成例を示す図である。要約処理装置６は、機械学習法により、文章の内容を示す重要文を抽出してその文章の要約を生成する処理装置であって、コーパス入力手段６１、タグ付与手段６２、コーパス記憶手段６３、ユーザ範囲抽出手段６４、教師データ変換手段６５、素性抽出手段６６、機械学習手段６７、学習結果記憶手段６８、データ入力手段６９、素性抽出手段６１０、要約推定手段６１１、タグ付与手段６１２、要約出力処理手段６１３、表示装置６１５、および入力装置６１６で構成される。

【０１５７】

要約処理装置６のタグ付与手段６２は、予め、重要文タグと範囲指定タグとを備えておく。重要文タグは、ある要約において重要な文である範囲を示す属性情報である。範囲指定タグは、ユーザが指定した重要文が要約する対象となる文の範囲を示す属性情報である。これらのタグはＳＧＭＬ形式であり、重要文タグは＜IMP __SENT＞＜/IMP__SENT＞と記述され、範囲指定タグは＜UC＞＜/UC ＞と記述される。

【０１５８】

コーパス入力手段６１は、コーパス７を入力する処理手段である。入力されるコーパス７は、テキストデータであって、例えば電子化された大量の新聞記事データ、論文データなどである。

【０１５９】

タグ付与手段６２は、入力されたコーパス７のテキストデータを表示装置６１５に表示し、表示装置６１５に表示されたテキストデータ上でユーザが重要文として指定した文の前後に重要文タグを付加し、さらに、表示されたテキストデータ上でユーザが指定した重要文が要約する対象となる文の範囲の前後に範囲指定タグを付加し、重要文タグおよび範囲指定タグが付与されたテキストデータを含むタグ付きコーパスをコーパス記憶手段６３に格納する処理手段である。

【０１６０】

または、タグ付与手段６２は、ユーザが重要文による要約の対象となる範囲を指定しなかった場合に、重要文タグが付与された文が含まれる所定の範囲を、ユーザが選択した要約の対象となる範囲とみなして、その範囲の前後に範囲指定タグを付加する。所定の範囲は、重要文タグが付与された文を含む段落、またはその文から前後所定数の文の範囲などの予め定めた規則をもとに決定する。

【０１６１】

ユーザ範囲抽出手段６４は、コーパス記憶手段６３に記憶されたタグ付きコーパスから教師データを作成するため、範囲指定タグで囲まれたユーザ範囲データ（段落データ）を抽出する手段である。

【０１６２】

教師データ変換手段６５は、抽出された段落データを文単位に分割し、抽出された段落データから重要文タグで囲まれた文（重要文）を検出し、分割した文のうち検出した重要文に重要文であることを示す分類ラベルを付与し、分割した文のうち検出した重要文以外の文に重要文でないことを示す分類ラベルを付与し、各文を教師データとする処理手段である。

【０１６３】

素性抽出手段６６は、段落データに対して形態素解析処理、構文解析処理などを行って所定の素性を抽出し、文ごとの素性の集合と付与された分類ラベルとの組を生成する手段である。

【０１６４】

素性として、例えば、１）文のなめらかさを示す情報、２）内容をよく表しているかどうかを示す情報、３）自動要約で用いられる情報などを抽出する。１）文のなめらかさを示す情報として、ｋ－ｇｒａｍ 形態素列のコーパスでの存在、かかりうけ文節間の意味的

整合度などを、また、２）内容をよく表しているかどうかを示す情報として、要約前のテキストにあったキーフレーズの包含率などを、また、３）自動要約で用いられる情報として、その文の位置やリード文かどうか、ＴＦ／ＩＤＦ（ＴＦは文書中でのその語の出現回数もしくは頻度を示す値、ＩＤＦはあらかじめ持っている多数の文書群のうち、その語が出現する文書数の逆数をいう）、文の長さ、固有表現・接続詞・機能語などの手がかり表現の存在などの情報を抽出する。

【０１６５】

機械学習手段６７は、分割した各文の素性の集合と分類ラベルとの組を利用して、各文について、その素性と集合との場合にどのような分類ラベルになりやすいかを学習し、学習結果を学習結果記憶手段６８に記憶する処理手段である。

【０１６６】

データ入力手段６９は、要約を行う文章、段落などのテキストデータを入力する処理手段である。

【０１６７】

素性抽出手段６１０は、入力データに対して形態素解析処理、構文解析処理などを行い、文ごとに所定の種類の素性を抽出する処理手段である。

【０１６８】

要約推定手段６１１は、学習結果記憶手段６８に記憶された学習結果を利用して、入力データの各文について、その素性の集合の場合になりやすい分類ラベルを推定する処理手段である。

【０１６９】

タグ付与手段６１２は、入力データ中の推定解に対応する文の前後に、重要文タグを挿入する処理手段である。

【０１７０】

要約出力処理手段６１３は、重要文タグで囲まれた文を要約として表示装置６１５に出力する処理手段である。

【０１７１】

表示装置６１５は表示装置２１、４９と、入力装置６１６は入力装置２２と、それぞれ同様の装置である。

【０１７２】

図１６および図１７は、要約処理の処理フローを示す図である。

【０１７３】

要約処理装置６のタグ付与手段６２は、ユーザが、ある文の要約として重要であると考える文（重要文）を示す分類ラベル「重要文」に対応する重要文タグ＜IMP __SENT＞＜/IMP__SENT＞と、ユーザが指定した重要文による要約の対象となる文の範囲を示す範囲指定タグ＜UC＞＜/UC ＞を用意しておく。

【０１７４】

要約処理装置６のコーパス入力手段６１が、テキストデータで構成されるコーパス７を入力する（ステップＳ３０）。コーパス７には、以下のようなテキストデータが含まれていたとする。

「…さらに、名詞の修飾語や所有者の情報を用い、より確実に指示対象の推定を行う。この結果、学習サンプルにおいて適合率82%、再現率85%の精度で、テストサンプルにおいて適合率79%、再現率77%の精度で、照応する名詞の指示対象の推定をすることができた。また、対照実験を行って名詞の指示性や修飾語や所有者を用いることが有効であることを示した。…」

タグ付与手段６２は、コーパス７のテキストデータを表示しユーザにタグ付与操作を促すタグ付与画面２００を表示装置６１５に表示する（ステップＳ３１）。

【０１７５】

図１８は、タグ付与画面２００の例を示す図である。タグ付与画面２００は、コーパス７のテキストデータを表示して分類タグを付加する箇所を指定できる指定項目２０１、タ

グ付与手段６２が備える分類先を選択できる選択項目２０３などで構成される。

【０１７６】

タグ付与画面２００でユーザによって重要文が指定されたら（ステップＳ３２）、タグ付与手段６２は、タグ付与画面２００で指定された文に対応する文字列の前後に選択された重要文タグ＜IMP _SENT＞＜/IMP_SENT＞を挿入する（ステップＳ３３）。さらに、ユーザによって、選択した重要文により要約される範囲が指定されたら（ステップＳ３４）、指定された範囲に対応するテキストデータの文字列の前後に範囲指定タグ＜UC＞＜/UC＞を付加する（ステップＳ３５）。

【０１７７】

例えば、図１８（Ａ）に示すように、ユーザが、指定項目２０１のテキストデータの以下の文をマウスドラッグ操作などにより、重要文として選択し、マウス右ボタンクリック操作などにより、選択項目２０３から重要文を選択する。

「この結果、学習サンプルにおいて、適合率82%、再現率85%の精度で、テストサンプルにおいて適合率79%、再現率77%の精度で、照応する名詞の指示対象の推定をすることができた。」

タグ付与手段６２は、以下のように、タグ付与画面２００で指定された文に対応する文字列の前後に選択された重要文タグ＜IMP _SENT＞＜/IMP_SENT＞を挿入する。

「＜IMP _SENT＞この結果、学習サンプルにおいて、適合率82%、再現率85%の精度で、テストサンプルにおいて適合率79%、再現率77%の精度で、照応する名詞の指示対象の推定をすることができた。＜/IMP_SENT＞」

さらに、ユーザが、以下の範囲を要約の対象とする範囲として指定する。例えば、図１８（Ｂ）に示すように、ユーザが、指定項目２０１のテキストデータの以下の文をマウスドラッグ操作などにより、ユーザ範囲データ（段落データ）として選択し、マウス右ボタンクリック操作などにより、選択項目２０３からユーザ範囲を選択する。

「さらに、名詞の修飾語や所有者の情報を用い、より確実に指示対象の推定を行う。…
この結果、学習サンプルにおいて適合率82%、再現率85%の精度で、テストサンプルにおいて適合率79%、再現率77%の精度で、照応する名詞の指示対象の推定をすることができた。また、対照実験を行って名詞の指示性や修飾語や所有者を用いることが有効であることを示した。」

タグ付与手段６２は、以下のように、指定された範囲に対応するテキストデータの文字列の前後に範囲指定タグ＜UC＞＜/UC ＞を付加する。

「＜UC＞さらに、名詞の修飾語や所有者の情報を用い、より確実に指示対象の推定を行う。…
＜IMP _SENT＞この結果、学習サンプルにおいて、適合率82%、再現率85%の精度で、テストサンプルにおいて適合率79%、再現率77%の精度で、照応する名詞の指示対象の推定をすることができた。＜/IMP_SENT＞また、対照実験を行って名詞の指示性や修飾語や所有者を用いることが有効であることを示した。＜/UC ＞」

なお、タグ付与手段６２は、ユーザが重要文により要約される範囲を指定しなかった場合には、重要文タグが付与された文を含む所定の範囲、例えば、重要文タグが付与された文を含む段落の範囲をユーザが指定した範囲とみなして、その範囲の前後に範囲指定タグを付加する(ステップＳ３６）。範囲の指定とみなす法としては、予め、同一文、同一の段落、前後所定数の文などと決めておく。

【０１７８】

タグ付与手段６２は、重要文タグおよび範囲指定タグが付与されたテキストデータを含むタグ付きコーパスの全部または一部をコーパス記憶手段６３に記憶する（ステップＳ３７）。

【０１７９】

そして、ユーザ範囲抽出手段６４は、コーパス記憶手段６３に記憶されたタグ付きコーパスから範囲指定タグで囲まれた範囲のテキストデータ（段落データ）を抽出する（ステップＳ３８）。さらに、教師データ変換手段６５は、抽出された段落データを文単位で分

割し、抽出された範囲中の重要文タグで囲まれた文（重要文）に分類ラベル「重要文」を付与して教師データとする（ステップＳ３９）。

【０１８０】

素性抽出手段６６は、教師データに対して所定の解析処理を行い、所定の素性を抽出する（ステップＳ３１０）。解析処理として、形態素解析、構文解析などを既知の処理手法を用いて行う。解析処理により、素性として、ｋ－ｇｒａｍ形態素列のコーパスでの存在、かかりうけ文節間の意味的整合度、テキストデータに存在するキーフレーズの包含率、重要文の位置、重要文がリード文かどうか、ＴＦ／ＩＤＦ、重要文の長さ、固有表現・接続詞・機能語などの手がかり表現の存在などの情報を抽出する。

【０１８１】

さらに、機械学習手段６７は、各文の素性の集合と分類ラベルとの組を利用して、各文について、素性の集合の場合にどのような分類先になりやすいかを学習し（ステップＳ３１１）、学習結果を学習結果記憶手段６８に記憶する（ステップＳ３１２）。

【０１８２】

データ入力手段６９は、要約対象の文章、段落などのテキストデータを入力する（ステップＳ３１３）。素性抽出手段６１０は、入力データに対して形態素解析処理、構文解析処理などを行い、所定の種類の素性を抽出する（ステップＳ３１４）。要約推定手段６１１は、学習結果記憶手段６８に記憶された学習結果を利用して、入力データの各文について、その素性の集合の場合になりやすい分類ラベルを推定する（ステップＳ３１５）。

【０１８３】

その後、タグ付与手段６１２は、入力データ中の推定された分類ラベルに対応する文字列の前後に重要文タグ＜IMP __SENT＞＜/IMP__SENT＞を挿入する（ステップＳ３１６）。要約出力処理手段６１３は、入力データ中の重要文タグで囲まれた文を要約として抽出し、表示装置６１５に表示する（ステップＳ３１７）。

【０１８４】

このように、ユーザが指定した重要文とその重要文により要約される文章との関連を機械学習により学習し、ユーザの指向に適応した要約を行うことができる。

【０１８５】

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。

【０１８６】

例えば、第１の実施例において、タグ登録手段１２により、範囲指定タグとして、ある特定の分類先についてユーザが付与作業を確認した分類先であることを示すタグ（ユーザ指定分類タグ用範囲指定タグ）を定義できるようにする。ユーザ指定分類タグ用範囲指定タグは、例えば＜UC-LOCATION ＞＜/UC-LOCATION＞と記述する。既に説明した処理により、ユーザは、教師データとする以下のような文を数多く生成したいとする。

「＜UC＞＜LOCATION＞日本＜/LOCATION ＞の首相は＜PERSON＞小泉＜/PERSON ＞さんです。＜/UC ＞小泉さんはいつも思いきったことをしています。」

しかし、ユーザが、分類先の「人名」は多く指定したが、「地名」はあまり多く指定していなかった場合には、分類先「地名」の指定だけをさらに増やしたいと考えることがある。

【０１８７】

このような場合に、ユーザは、まず、通常どおり分類先「地名」に対応する分類タグ＜LOCATION＞＜/LOCATION ＞を付与する単語を指定し、さらにユーザ指定分類タグ用範囲指定タグを使用して、タグ付与作業をチェックした範囲を指定する。

【０１８８】

タグ付与手段１４は、これらの指定をもとに、以下のようにテキストデータに分類タグおよびユーザ指定分類タグ用範囲指定タグを挿入する。

「＜UC-LOCATION ＞＜LOCATION> 大阪＜/LOCATION ＞の知事は太田さんです。＜/UC-LOCATION＞大阪は古くは商業の中心地でした。」

そして、以降の処理においては分類先「地名」についてのみ処理を行うようにする。

【０１８９】

これにより、ユーザのタグ付与作業のチェックがより効率的になり、教師データの量が不足していたような分類先だけを重点的に増やすことが可能となる。

【０１９０】

また、コーパス入力手段１１で、コーパス２として既にタグなどが付与されたコーパスを入力し、タグ付与手段１４では、既に付与されたタグを削除し、新しく分類タグを付けなおすことにより、タグの修正をするようにしてもよい。

【０１９１】

この場合に、タグ付与手段１４は、予め修正確認範囲タグを用意しておき、入力されたコーパス２の一部で分類タグを完全に修正した範囲としてユーザが指定した範囲の前後に、この修正確認範囲タグを挿入し、ユーザ範囲抽出手段１６では、この修正確認範囲タグで指定された範囲をユーザ範囲として抽出するようにしてもよい。これにより、より精度の高い教師データを作成することが可能となる。

【０１９２】

また、第１の実施例では、機械学習法として、サポートベクトルマシン法を用いる場合の処理例を説明したが、これ以外に、決定リスト法、最大エントロピー法などの教師データを用いた機械学習法を用いた処理を行ってもよい。

【０１９３】

決定リスト法は、あらかじめ設定しておいた素性f_i（∈Ｆ，１≦ｊ≦ｋ）のうちいずれか１つのみを文脈として各分類の確率値を求め、その確率値が最も大きい分類を求める分類とする方法である。ある文脈ｂで分類ａを出力する確率は、以下の式によって与えられる。

【０１９４】

【数３】

$$p(a|b) = \tilde{p}(a|f_{max}) \qquad (7)$$

ただし，

$$f_{max} = argmax_{f_j \in F}\ max_{a_i \in A}\ \tilde{p}(a_i|f_j) \qquad (8)$$

また、［ｐ～］（a_i｜f_i）（［ｐ～］は、ｐチルダ（～）を示す）は、学習データで素性f_jを文脈に持つ場合の分類a_iの出現の割合である。

【０１９５】

言語解析処理装置４が、決定リスト法を用いる場合には、機械学習手段４２では、素性の集合と分類先との対で構成したものを規則とし、前記規則を所定の順序でリスト上に並べたものを学習結果とし、学習結果として得られたその規則のリストを学習結果記憶手段４３に記憶する。そして、解推定手段４６では、学習結果記憶手段４３に記憶されている規則のリストを参照して、リストを先頭からチェックしていき、入力データの素性の集合と一致する規則を探し出し、その規則の分類先を、その素性の集合のときになりやすい分類先として推定する。

【０１９６】

最大エントロピー法は、予め設定しておいた素性f_j（１≦ｊ≦ｋ）の集合をＦとするとき、式（９）を満足しながらエントロピーを意味する式（１０）を最大にするときの確率分布ｐ（ａ，ｂ）を求め、その確率分布に従って求まる各分類の確率のうち、最も大きい確率値を持つ分類を求める方法である。

【０１９７】

【数４】

$$\sum_{a\in A,b\in B} p(a,b)g_j(a,b) = \sum_{a\in A,b\in B} \tilde{p}(a,b)g_j(a,b) \quad (9)$$
$$for\ \forall f_j\ (1 \le j \le k)$$

$$H(p) = -\sum_{a\in A,b\in B} p(a,b)\ log\,(p(a,b)) \quad (10)$$

ただし、Ａ、Ｂは分類と文脈との集合を意味する。g_j（ａ，ｂ）は、文脈ｂに素性f_jがあって、かつ分類がａの場合に１となり、それ以外で０となる関数を意味する。また、［ｐ～］（ａ，ｂ）は既知のデータでの（ａ，ｂ）の出現の割合を意味する。

【０１９８】

式（９）は、確率ｐと出力と素性との組の出現を意味する関数ｇをかけることで出力と素性の組の頻度の期待値を求めることになっており、右辺の既知データにおける期待値と、左辺に求める確率分布にもとづいて計算される期待値が等しいことを制約として、エントロピーの最大化（確率分布の平滑化）を行って、出力と文脈の確率分類を求めるものである。

【０１９９】

言語解析処理装置４が、最大エントロピー法を用いる場合には、機械学習手段４２では、教師データから解となりうる分類先を特定し、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素性の集合と解となりうる分類先の二項からなる確率分布を求め、前記確率分布を学習結果とし、その確率分布を学習結果記憶手段４３に記憶する。そして解推定手段４６では、学習結果記憶手段４３に記憶されている学習結果の確率分布にもとづいて、入力データの素性の集合の場合のそれぞれの解となりうる分類先の確率を求め、最も大きい確率値を持つ解となりうる分類先を特定し、その特定した分類先を入力データの素性の集合の場合になりやすい分類先と推定する。

【０２００】

また、言語解析処理装置４が、サポートベクトルマシン法を用いる場合には、機械学習手段４２では、教師データから解となりうる分類先を特定し、その分類先を正例と負例に分割し、所定のカーネル関数を用いたサポートベクトルマシン法を実行する関数にしたがって素性の集合を次元とする空間上で正例と負例の間隔を最大にして正例と負例を超平面で分割する超平面を求め、その超平面を学習結果とし、その超平面を学習結果記憶手段４３に記憶する。そして、解推定手段４６では、学習結果記憶手段４３に記憶されている学習結果の超平面を利用して、入力データの素性の集合がこの超平面で分割された空間において正例側か負例側のどちらにあるかを特定し、その特定された結果に基づいて定まる分類先を、入力データの素性の集合の場合になりやすい分類先と推定する。

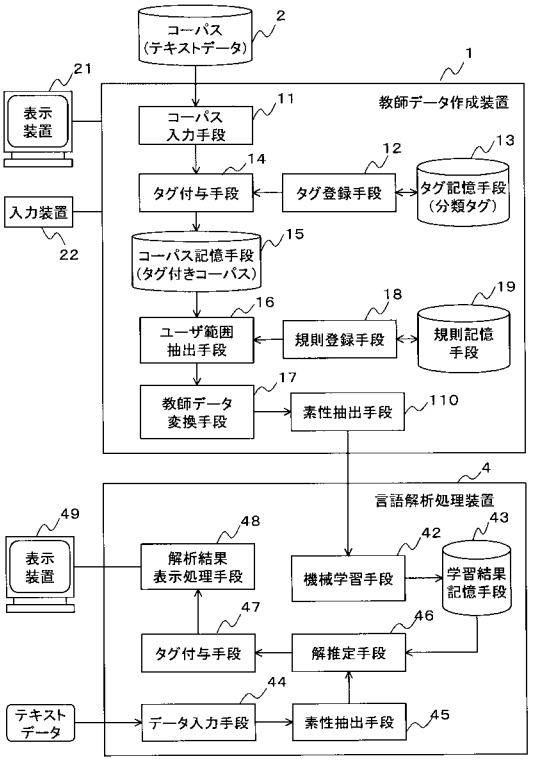
【図面の簡単な説明】

【０２０１】

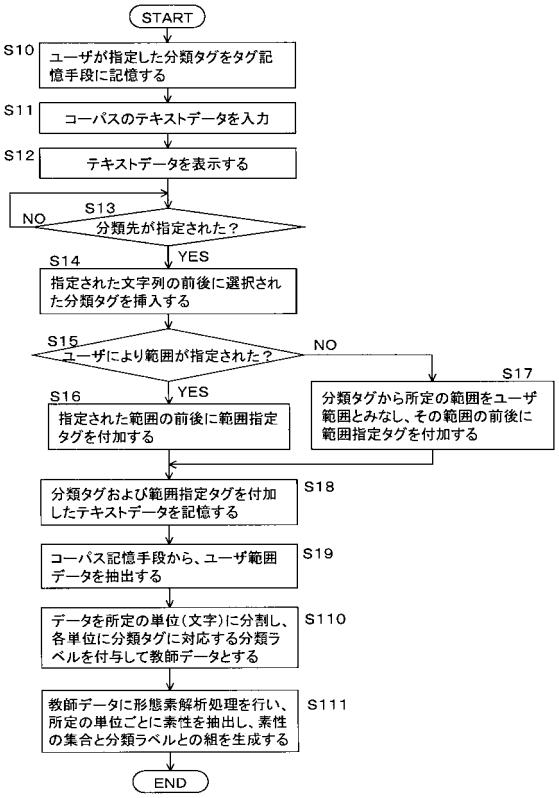
【図１】機械学習法を用いた言語解析処理を行う場合の本発明の構成例を示す図である。
【図２】教師データ作成処理の処理フローを示す図である。
【図３】タグ付与画面の例を示す図である。
【図４】教師データの例を示す図である。
【図５】教師データの各文字の素性と分類ラベルとの組の例を示す図である。
【図６】言語解析処理の処理フローを示す図である。
【図７】サポートベクトルマシン法の最大マージンを説明するための図である。
【図８】入力データとその各文字に付与された分類ラベルの例を示す図である。
【図９】タグ付与画面の例を示す図である。
【図１０】教師データの例を示す図である。

【図１１】分類タグが付与された入力データの例を示す図である。
【図１２】教師データの例を示す図である。
【図１３】テキストデータの各単語の素性と分類ラベルとの組の例を示す図である。
【図１４】テキストデータの各単語の素性と分類ラベルとの組の例を示す図である。
【図１５】機械学習法による要約処理を行う場合の本発明の構成例を示す図である。
【図１６】要約処理の処理フローを示す図である。
【図１７】要約処理の処理フローを示す図である。
【図１８】タグ付与画面の例を示す図である。
【符号の説明】
【０２０２】
１　教師データ作成装置
１１　コーパス入力手段
１２　タグ登録手段
１３　タグ記憶手段
１４　タグ付与手段
１５　コーパス記憶手段
１６　ユーザ範囲抽出手段
１７　教師データ変換手段
１８　規則登録手段
１９　規則記憶手段
１１０　素性抽出手段
２１　表示装置
２２　入力装置
２　コーパス（テキストデータ）
４　言語解析処理装置
４２　機械学習手段
４３　学習結果記憶手段
４４　データ入力手段
４５　素性抽出手段
４６　解推定手段
４７　タグ付与手段
４８　解析結果表示処理手段
４９　表示装置
６　要約処理装置
６１　コーパス入力手段
６２　タグ付与手段
６３　コーパス記憶手段
６４　ユーザ範囲抽出手段
６５　教師データ変換手段
６６　素性抽出手段
６７　機械学習手段
６８　学習結果記憶手段
６９　データ入力手段
６１０　素性抽出手段
６１１　要約推定手段
６１２　タグ付与手段
６１３　要約出力処理手段
６１５　表示装置
６１６　入力装置
７　コーパス（テキストデータ）

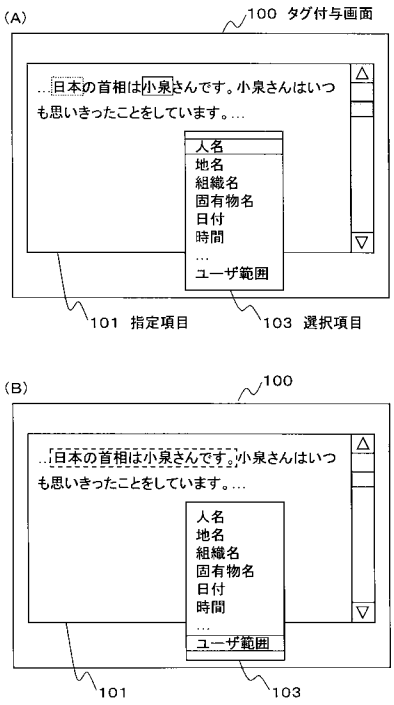
【図１】



【図２】



【図３】



【図４】

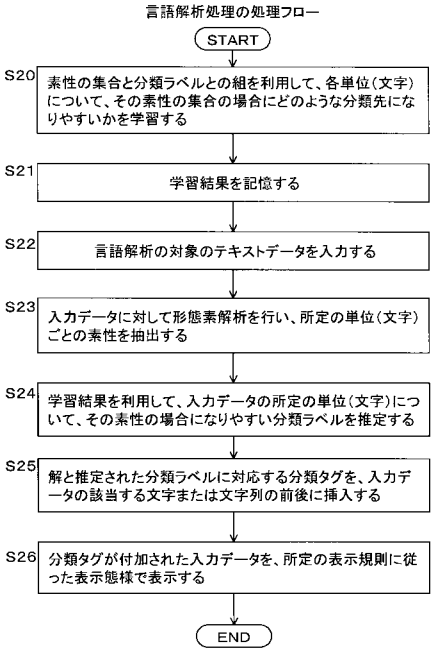
教師データの例

日	B-LOCATION
本	I-LOCATION
の	O
首	O
相	O
は	O
小	B-PERSON
泉	I-PERSON
さ	O
ん	O
で	O
す	O
。	O

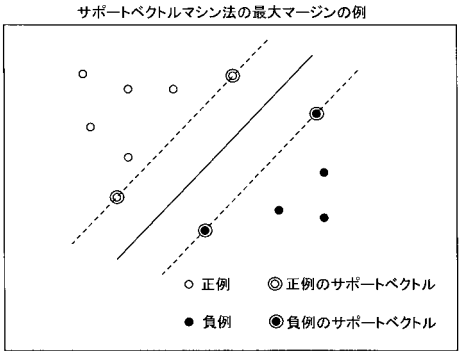
【図５】

組:			
	日	日本、名詞－固有、B	B-LOCATION
	本	日本、名詞－固有、I	I-LOCATION
	の	の、助詞－格助詞、S	O
	首	首相、名詞－一般、B	O
	相	首相、名詞－一般、I	O
	は	は、助詞－格助詞、S	O
→	小	小泉、名詞－固有、B	B-PERSON
	泉	小泉、名詞－固有、I	I-PERSON
	さ	さん、名詞－接尾、B	O
	ん	さん、名詞－接尾、I	O
	で	です、助動詞－形動、B	O
	す	です、助動詞－形動、I	O
	。	読点	O

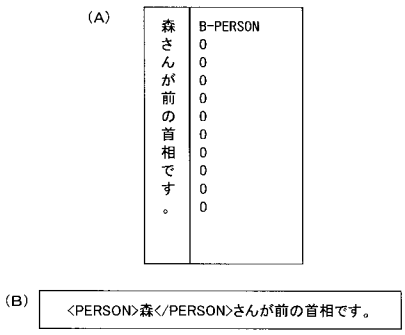
【図６】



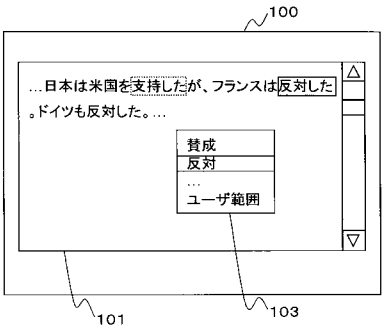
【図７】



【図８】



【図９】



【図１０】

日	O
本	O
は	O
米	O
国	O
を	O
支	B-APPROVAL
持	I-APPROVAL
し	I-APPROVAL
た	I-APPROVAL
が	O
、	O
フ	O
ラ	O
ン	O
ス	O
は	O
反	B-DISAPPROVAL
対	I-DISAPPROVAL
し	I-DISAPPROVAL
た	I-DISAPPROVAL
。	O

【図１１】

ロシアは<DISAPPROVAL>反対した</DISAPPROVAL>。

【図１２】

教師データの例

ref0	おじいさん	個体導入
ref0	おじいさん	おじいさん〔１番目〕を指示
ref1	山	個体導入
ref1	そこ	山〔１番目〕を指示

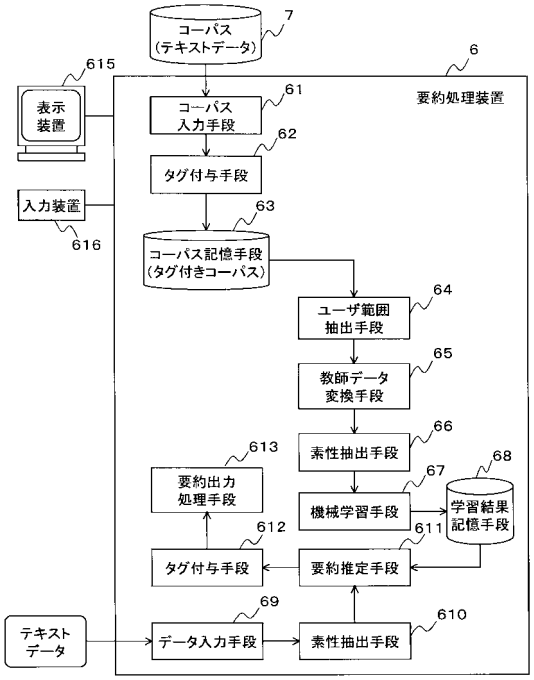
【図１３】

おじいさん	おじいさん，個体導入，０個，個体導入，個体導入，なし	個体導入
おじいさん	おじいさん，個体導入，０個，個体導入，個体導入，あり	１番目指示
山	山，個体導入，０個，個体導入，個体導入，なし	個体導入
そこ	そこ，個体導入，０個，個体導入，個体導入，なし	１番目指示

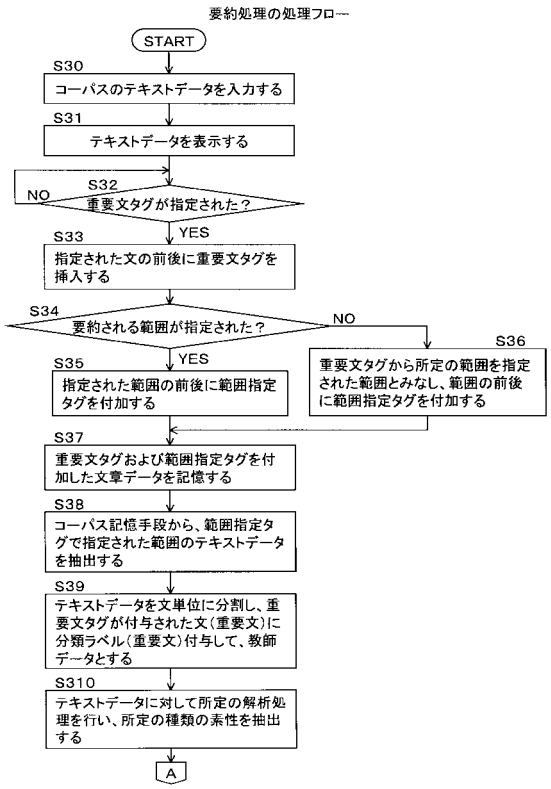
【図１４】

おじいさん		個体導入
おじいさん	おじいさん，おじいさん，２個，整合性有，整合性有，あり	１番目指示
山		個体導入
そこ	そこ，山，２個，整合性有，整合性有，なし	１番目指示

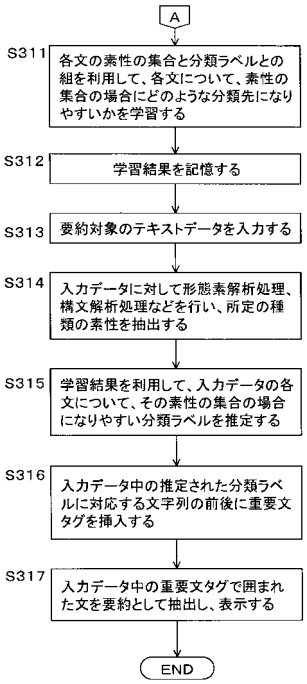
【図１５】



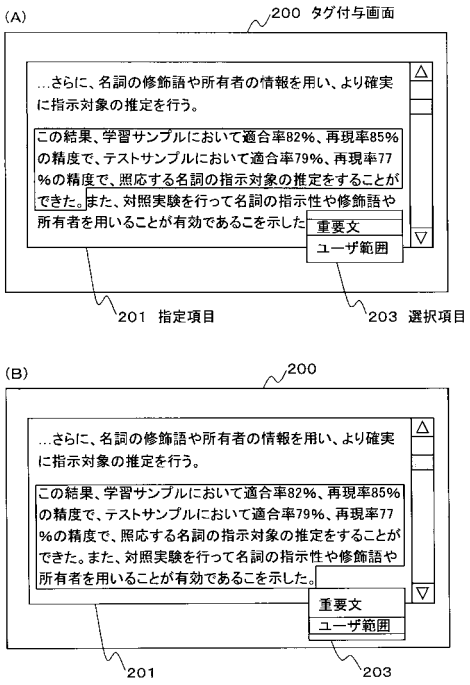
【図１６】



【図１７】



【図１８】


(A)
200 タグ付与画面
…さらに、名詞の修飾語や所有者の情報を用い、より確実に指示対象の推定を行う。
この結果、学習サンプルにおいて適合率82%、再現率85%の精度で、テストサンプルにおいて適合率79%、再現率77%の精度で、照応する名詞の指示対象の推定をすることができた。また、対照実験を行って名詞の指示性や修飾語や所有者を用いることが有効であることを示した。
重要文
ユーザ範囲
201 指定項目
203 選択項目
(B)
200
…さらに、名詞の修飾語や所有者の情報を用い、より確実に指示対象の推定を行う。
この結果、学習サンプルにおいて適合率82%、再現率85%の精度で、テストサンプルにおいて適合率79%、再現率77%の精度で、照応する名詞の指示対象の推定をすることができた。また、対照実験を行って名詞の指示性や修飾語や所有者を用いることが有効であることを示した。
重要文
ユーザ範囲
201
203

フロントページの続き

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １７／２７－１７／２８