

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2016年7月28日(28.07.2016)



(10) 国際公開番号

WO 2016/117607 A1

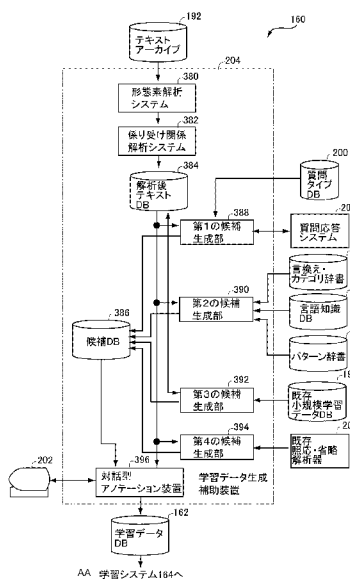
- (51) 国際特許分類:
G06F 17/27 (2006.01) G06F 17/24 (2006.01)
- (21) 国際出願番号: PCT/JP2016/051577
- (22) 国際出願日: 2016年1月20日(20.01.2016)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2015-011491 2015年1月23日(23.01.2015) JP
- (71) 出願人: 国立研究開発法人情報通信研究機構
(NATIONAL INSTITUTE OF INFORMATION AND COMMUNICATIONS TECHNOLOGY) [JP/JP]; 〒1848795 東京都小金井市貫井北町4-2-1 Tokyo (JP).

- (72) 発明者: 飯田 龍(IIDA, Ryu); 〒1848795 東京都小金井市貫井北町4-2-1 国立研究開発法人情報通信研究機構内 Tokyo (JP). 鳥澤 健太郎(TORISAWA, Kentaro); 〒1848795 東京都小金井市貫井北町4-2-1 国立研究開発法人情報通信研究機構内 Tokyo (JP). 橋本 力(HASHIMOTO, Chikara); 〒1848795 東京都小金井市貫井北町4-2-1 国立研究開発法人情報通信研究機構内 Tokyo (JP). 呉 鍾勲(OH, Jonghoon); 〒1848795 東京都小金井市貫井北町4-2-1 国立研究開発法人情報通信研究機構内 Tokyo (JP). 大竹 清敬(OTAKE, Kiyonori); 〒1848795 東京都小金井市貫井北町4-2-1 国立研究開発法人情報通信研究機構内 Tokyo (JP). 木俣 豊(KIDAWARA, Yutaka); 〒1848795 東京都小金井市貫井北町4-2-1 国立研究開発法人情報通信研究機構内 Tokyo (JP).

[続葉有]

(54) Title: ANNOTATION ASSISTANCE DEVICE AND COMPUTER PROGRAM THEREFOR

(54) 発明の名称: アノテーション補助装置及びそのためのコンピュータプログラム



162 Learning data database
190 Existing small-scale learning data database
192 Text archive
194 Language knowledge database
196 Paraphrase/category dictionary
198 Pattern dictionary
200 Question type database
204 Learning data generation assistance device
206 Question response system
208 Existing anaphor/abbreviation analyzer
380 Morphological analysis system
382 Dependency relationship analysis system
384 Post-analysis text database
386 Candidate database
388 First candidate generation unit
390 Second candidate generation unit
392 Third candidate generation unit
394 Fourth candidate generation unit
396 Interactive annotation device
AA To learning system 164

(57) Abstract: [Problem] To provide an annotation assistance device with which annotations with respect to a text can be constructed easily and inexpensively. [Solution] An annotation data generation assistance system 160 includes: an input/output device 202 that receives interactive processing inputs; a morphological analysis system 380 and a dependency relationship analysis system 382 that perform a morphological analysis and a dependency relationship analysis on text data in a text archive 192; first through fourth candidate generation units 388, 390, 392, and 394 that detect abbreviations and reference terms in the dependency relationships of predicates in morpheme strings, identify a position to be subjected to an annotation process, and use language knowledge to suggest candidate expressions to be inserted; a candidate database (DB) 386 that stores the suggested candidates; and an interactive annotation device 396 that reads annotation process candidates from the candidate DB 386 and appends, as an annotation, a candidate selected in an interactive process by the input/output device 202.

(57) 要約: 【課題】テキストへのアノテーションを容易に、低コストで構築できるアノテーション補助装置を提供する。【解決手段】アノテーションデータ生成補助システム160は、対話型処理で入力を受ける入出力装置202と、テキストアーカイブ192のテキストデータに形態素解析及び係り受け解析を行なう形態素解析システム380及び係り受け関係解析システム382と、形態素列中の述語の係り受け関係において、省略又は指示語を検出し、アノテーション処理の対象位置を特定し、挿入される表現の候補を言語知識を用いて推定する第1～第4の候補生成部388、390、392及び394と、推定された候補を記憶する候補DB386と、アノテーション処理の候補を候補DB386から読出して、入出力装置202による対話型処理で選択された候補をアノテーションとして付加する対話型アノテーション装置396を含む。



(74) 代理人: 清水 敏 (SHIMIZU, Satoshi); 〒5300002
大阪府大阪市北区曽根崎新地二丁目5番3号
堂島TSSビル4階 Osaka (JP).

(81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーロパ (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告 (条約第 21 条(3))

明 細 書

発明の名称：

アノテーション補助装置及びそのためのコンピュータプログラム

技術分野

[0001] この発明は、自然言語処理技術に関し、特に、自然言語の文章中に頻出する省略、代名詞等の指示語の照応関係についてのアノテーションを、テキストに対して効率的に行うためのアノテーション補助装置に関する。

背景技術

[0002] 自然言語のテキスト中には、省略及び指示語が頻出する。例えば図1に示す例文30を考える。例文30は第1文と第2文とからなる。第2文には、「それ」という指示語（代名詞）42が含まれる。この場合、「それ」という指示語は第1文の「モン歴の正月の日付」という表現40を指す。このように指示語の指す語を特定する処理を「照応解析」と呼ぶ。一方、図2の例文60を考える。この例文60は、第1文と第2文とからなる。第2文において、「自己診断機能を搭載」という述部の主語は省略されているが、この主語の省略箇所76には、第1文の「新型交換機」という単語72が省略されている。同様に、「200システムを設置する予定だ。」という述部の主語は省略されている。この主語の省略箇所74には、第1文の「N社」という単語70が省略されている。このように、主語等の省略を検出し、それを補完する処理を「省略解析」と呼ぶ。以後、照応解析と省略解析とをまとめて「照応・省略解析」と呼ぶ。

[0003] いわゆる人工知能において、人間とのコミュニケーションをとるために、自然言語処理は欠かせない技術である。自然言語処理の重要な問題として、自動翻訳及び質問応答等が存在する。照応・省略解析の技術は、このような自動翻訳及び質問応答において必須の要素技術である。しかし、現状の照応・省略解析の技術水準は、言語にもよるが未だ実用レベルには至っていない。

[0004] このように照応・省略解析の性能が向上しない理由は種々考えられる。主な理由として、このような解析には一般常識が必要であることが挙げられる。一般常識を考慮するためには、照応・解析に関する人間による判断を考慮する必要がある。しかし、そうした要素を考慮した解析アルゴリズムは難しい。結局、人間による判断を学習データとして多数準備し、統計的学習により照応・省略解析を行う解析器を作成する必要がある。しかし、そのような解析器構築用の学習データを作成するためのコストは非常に大きいことが知られている。そのため、十分な量の学習データが作成できないために照応・省略解析の性能の向上を図ることができない。

[0005] 自然言語処理の性能を高め、ひいては人工知能をより知的にするためには、照応・省略解析に関するこうした問題を解決する事が必要である。

[0006] 照応・省略解析器の学習データを作成する技術として、後掲の非特許文献1に記載されたものがある。この技術では、学習データの作成は以下のように行われる。すなわち、人間がテキストを最初から読み、代名詞及び省略等を人手で検出するという作業が行われる。機械による支援としては、例えば照応関係の指し先の表現を予め列挙する程度である。

[0007] 一方、後掲の非特許文献2には、自動的に、一定の手順で照応・省略解析を行う照応・省略解析アルゴリズムが開示されている。この文献に開示された技術は、形態素解析、構文・係り受け解析（以下単に「係り受け解析」と呼ぶ。）等が出力する情報と、辞書等の外部資源ツールを利用する。ここで利用される辞書とは、例えば「食べる」という動詞の目的語には「食べ物」に相当する表現が含まれる、という特性をまとめたものである。非特許文献2に開示された技術では、これ以外にはテキストの中から得られる情報を参照して代名詞の指し先及び省略の補完が行われている。

先行技術文献

非特許文献

[0008] 非特許文献1：飯田龍、小町守、井之上直也、乾健太郎、松本裕治、述語構造と照応関係のアノテーション： NAIST テキストコーパス構築の経験から、

自然言語処理, Vol.17, No.2, pp.25-50, 2010.

非特許文献2: Ryu Iida, Massimo Poesio. A Cross-Lingual LLP Solution to Zero Anaphora Resolution. The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT2011), pp.804-813. 2011.

発明の概要

発明が解決しようとする課題

[0009] しかし、前掲の非特許文献2によれば、日本語の照応解析の場合で、再現率が44%、精度が42%、主語の省略解析の場合で再現率及び精度がいずれも35%という数字が報告されている。日常生活で目にする文書には、省略及び代名詞等が頻繁に使用されている。正確な資源言語処理を行うためには、省略箇所を検出、省略箇所及び代名詞の指示先の検出が正確に行われなければならない。そのためには、精度の高い照応・省略解析器を得る必要がある。しかし、前述のように、そうした照応・省略解析器の学習を行うためには、学習データの作成に手間がかかり、コストが掛かり過ぎるという問題があった。学習データは、文章中に出現する省略、代名詞等の指示語の照応関係についてのアノテーションを、テキストに対して行う事により作成する。もちろん、アノテーションの精度自体は高くしなければならないため、人間による判断を学習データに反映させなければならない。すなわち、アノテーション自体は人間が行うこととし、そうした人間のアノテーション作業をより効率的に行えるようなシステムを得ることが上記課題を解決する上で望ましい。

[0010] それゆえにこの発明の目的は、人間が、テキストに対するアノテーションデータを容易に構築でき、それによってコストを下げることのできるアノテーション補助装置を提供することである。

課題を解決するための手段

[0011] 本発明の第1の局面にかかるアノテーション補助装置は、自然言語のテキストの照応・省略解析のためのアノテーションデータの作成を補助する。こ

のアノテーション補助装置は、表示装置と入力装置とからなり、ユーザとの対話型処理によりユーザ入力を受け付ける入出力手段と、アノテーションの対象となるテキストアーカイブからテキストデータを読み出す読出手段と、読出手段が読み出したテキストデータの形態素解析及び係り受け解析を行って、係り受け構造を示す情報が付された形態素列を出力する解析手段と、解析手段の出力する形態素列中の述語を検索する述語検索手段と、述語検索手段が検索した述語の各々が関与する係り受け関係において、当該述語に対して所定の関係にあるべき単語が省略されていること、又は指示語となっていることを検出し、アノテーション処理の対象として当該単語の位置を特定するための対象特定手段と、対象特定手段が特定した単語の位置の各々について、当該位置に挿入されるべき表現の候補を、当該位置と周囲の形態素列との関係、及び予め準備された言語知識を用いて推定する候補推定手段と、候補推定手段により推定された候補を単語の位置と関連付けて記憶する候補記憶手段と、アノテーション処理の対象の各々について、候補推定手段が推定した候補を候補記憶手段から読出し、ユーザがいずれかを選択可能なように表示装置に表示する候補表示手段と、候補表示手段により表示された候補のいずれかを選択するユーザの指示に応答して、選択された候補を位置にアノテーションとして付加する対話的選択手段とを含む。

[0012] 好ましくは、候補推定手段は、対象特定手段が特定した単語の位置の各々について、当該単語の位置に入るべき単語を問い合わせる文を、当該位置の周囲の単語列を用いて生成する質問文生成手段と、質問文生成手段が生成した質問文を、予め準備された質問応答システムに入力として与える質問文入力手段と、質問文入力手段から与えられた質問文に対して質問応答システムから得られる回答から、単語の位置に挿入されるべき単語の候補を当該単語の位置と関連付けて候補記憶手段に記憶させる手段とを含む。

[0013] より好ましくは、候補推定手段は、さらに、複数の言語知識規則を記憶する言語知識記憶手段と、対象特定手段が特定した単語の位置の各々について、当該単語の位置と、当該単語の位置を特定する際に用いられた述語とを含

む表現であって、言語知識記憶手段に記憶された言語知識のいずれかに適合する表現を特定する適合表現抽出手段と、適合表現抽出手段により抽出された表現と、当該表現が適合した言語知識規則とを照合することにより、表現中の、対象特定手段が特定した単語の位置に挿入されるべき表現を、当該単語の位置に挿入されるべき文字列の候補として候補記憶手段に記憶させる手段とを含む。

[0014] さらに好ましくは、候補推定手段はさらに、予め準備された既存のアノテーション済テキストデータベースに含まれるテキスト中で、省略又は照応関係に関するアノテーションが付されている箇所を検索するアノテーション検索手段と、アノテーション検索手段により検索された箇所の各々について、各アノテーションごとに予め定められた方式にしたがって文を変形し、当該変形箇所を検索された箇所に対するアノテーションの候補として候補記憶手段に記憶させる手段とを含む。

[0015] 候補推定手段はさらに、解析手段の出力する解析後の形態素列を、他の既存の照応・省略解析器への入力として与える手段と、既存の照応・省略解析器による照応・省略解析結果を受け、当該照応・省略解析結果を、候補記憶手段に記憶させる手段とを含んでもよい。

[0016] 本発明の第2の局面にかかるコンピュータプログラムは、自然言語のテキストの照応・省略解析のためのアノテーションデータの作成を補助するアノテーション補助装置としてコンピュータを動作させる。このコンピュータプログラムは、コンピュータを、当該コンピュータの表示装置と入力装置とを用いたユーザとの対話型処理によりユーザ入力を受け付ける入出力手段と、アノテーションの対象となるテキストアーカイブからテキストデータを読み出す読出手段と、読出手段が読み出したテキストデータの形態素解析及び係り受け解析を行って、係り受け構造を示す情報が付された形態素列を出力する解析手段と、解析手段の出力する形態素列中の述語を検索する述語検索手段と、述語検索手段が検索した述語の各々が関与する係り受け関係において、当該述語に対して所定の関係にあるべき単語が省略されていること、又は

指示語となっていることを検出し、アノテーション処理の対象として当該単語の位置を特定するための対象特定手段と、対象特定手段が特定した単語の位置の各々について、当該位置に挿入されるべき表現の候補を、当該位置と周囲の形態素列との関係、及び予め準備された言語知識を用いて推定する候補推定手段と、候補推定手段により推定された候補を単語の位置と関連付けて記憶する候補記憶手段と、アノテーション処理の対象の各々について、候補推定手段が推定した候補を候補記憶手段から読出し、ユーザがいずれかを選択可能なように表示装置に表示する候補表示手段と、候補表示手段により表示された候補のいずれかを選択するユーザの指示に応答して、選択された候補を位置にアノテーションとして付加する対話的選択手段として機能させる。

図面の簡単な説明

[0017] [図1]照応解析を説明するための模式図である。

[図2]省略解析を説明するための模式図である。

[図3]省略解析のプロセスを説明するための模式図である。

[図4]本発明の一実施の形態に係る学習データ生成補助装置を含む照応・省略解析のためのシステムの全体ブロック図である。

[図5]言語知識を構成する規則の概略構成を説明するための図である。

[図6]質問タイプDBの概略構成を説明する模式図である。

[図7]質問応答システムを利用したアノテーション方法の概略を示す模式図である。

[図8]学習データ生成補助装置のブロック図である。

[図9]アノテーション候補を生成する第1の候補生成部のブロック図である。

[図10]アノテーション候補を生成する第2の候補生成部のブロック図である。

[図11]言語知識を利用したアノテーション候補の生成プロセスを説明する図である。

[図12]言語知識を利用したアノテーション候補の生成プロセスを説明する図

である。

[図13]言語知識を利用したアノテーション候補の生成プロセスを説明する図である。

[図14]言語知識を利用したアノテーション候補の生成プロセスを説明する図である。

[図15]アノテーション候補を生成する第3の候補生成部のブロック図である。

[図16]アノテーション候補を生成する第4の候補生成部のブロック図である。

[図17]テキストへのアノテーションをユーザとの対話により実行して学習データを生成するためのプログラムの状態遷移図である。

[図18]検出器学習装置222の詳細な機能ブロック図である。

[図19]学習データを生成するためのプログラムを実行するコンピュータの外観を示す図である。

[図20]図19に外観を示すコンピュータのハードウェアブロック図である。

発明を実施するための形態

[0018] 以下の説明及び図面では、同一の部品には同一の参照番号を付してある。したがって、それらについての詳細な説明は繰返さない。

[0019] [概略]

図3を参照して、本発明の第1の実施の形態に係る学習データ生成補助装置においてアノテーションをテキストに付する作業の概略を説明する。図3の上段を参照して、表示装置の画面に、アノテーション対象のテキスト90が表示される。このテキスト90は2つの文からなる。第1の文には、「導入する」という述部と、述部の主語である「N社」という単語100と、述部の目的語である「新型交換機」という単語110とが含まれる。第2の文は実質的には2つの節からなる。前半の文の先頭には「自己診断機能を搭載」の主語が省略されている箇所（省略箇所）114がある。前半の文と後半の文との中間箇所には、「200システムを設置する予定だ。」の主語が省

略されている省略箇所 104 がある。

[0020] 照応・省略解析では、矩形 102 及び矩形 112 内に記載したように、まず省略箇所 104 又は 114 を発見する。その後、その省略箇所を補完すべき表現を探索し決定する。省略箇所 104 の場合には単語 100 が補完すべき表現となる。省略箇所 114 の場合には、単語 110 が補完すべき表現となる。

[0021] 従来は、こうした省略箇所の発見と、補完表現の探索及び決定とを、いずれも人間が行っていた。しかしそのような作業は非効率的であり、学習データの生成コストを押し上げる要因となっていた。本実施の形態では、上記した省略箇所の発見と、補完表現の候補の発見とをコンピュータプログラムで自動的に行う。こうした処理には、後述するように様々な手法を使用する。特に、言語知識を活用することで、人間が行うような作業と同様、省略箇所の発見と補完表現の候補の探索とを効率的に行う。

[0022] 省略箇所の発見と補完表現の候補の発見とが終わると、各省略箇所に対して、図 3 の下段に示すような処理を行う。この例では、図 3 の上段に示す省略箇所 114 に対する補完表現に対しては「新型交換機」という単語 110 が発見され、省略箇所 104 には単語 100 を含む複数の単語が発見されたものとする。この場合、学習データ生成補助装置は、省略箇所 114 に単語 110 と同じ「新型交換機」という単語と主格の格助詞とからなる文字列 130 を挿入し、さらにその右側には、いわゆるドロップダウンリストが存在することを意味するアイコン「▼」を表示する。ユーザがこのアイコンをクリックするとドロップダウンリスト 132 が表示される。ドロップダウンリスト 132 には、表示された候補をこの省略箇所 114 に対するアノテーションとして付加して良いか否かを尋ねる 2 つのオプション（YES、NO）と、人間が手入力するオプション（「手入力」）が表示される。利用者が YES を選択すると、この省略箇所 114 に文字列 130 がアノテーションとして付加される。NO を選択すればこの省略箇所 114 にはアノテーションは付加されない。「手入力」を選択すると、図示しない入力ダイアログが開

き、アノテーションを直接入力できる。

[0023] 省略箇所 104 の場合、候補の一つ、例えば単語 100 と同じ単語と格助詞「ハ」からなる文字列 134 が表示され、右側にドロップダウンリストを示すアイコン「▼」が表示される。このアイコンをユーザがクリックすると、複数の候補のリストと「直接入力」とをオプションとして含むドロップダウンリストが表示される。ユーザが候補のいずれかを選択すると、その候補の単語が省略箇所 104 に対するアノテーションとして付加される。直接入力を選択されたときについては、省略箇所 114 と同様である。

[0024] [照応・解析システムの全体構成]

図 4 を参照して、最初に、本実施の形態に係る学習データ生成補助装置を使用する照応・省略解析システム 150 の全体構成について説明する。

[0025] この照応・省略解析システム 150 は、予め準備された大量のテキストについて、省略箇所及び照応詞の検出を行い、それらが指す表現（先行詞）の候補をユーザに提示して簡単な操作で選択させることにより、ユーザが照応・省略解析に必要なモデルの学習データ（アノテーションデータ）を生成するのを補助するアノテーションデータ生成補助システム 160 と、アノテーションデータ生成補助システム 160 を用いてユーザにより生成された学習データを記憶する学習データ DB 162 と、学習データ DB 162 に記憶された学習データを用いて、後述する省略検出器 166、照応詞検出器 168、及び先行詞特定器 170 の学習を行うための学習システム 164 とを含む。省略検出器 166 は、テキスト中の任意の述語を対象に、省略検出を行う。省略検出器 166 は、テキスト中の主語、目的語等が省略された省略箇所を検出するための統計的モデルを含む。照応詞検出器 168 は、テキスト中の任意の名詞句（照応詞候補）を分類対象として、他の単語などを指示するための照応詞を検出するための統計的モデルを含む。先行詞特定器 170 は、テキスト中の、省略を持つ述語と先行詞候補である名詞句とを分類対象に、省略が指す先（先行詞）を特定するための検出器と、テキスト中の名詞句の対（照応詞と先行詞候補）とを分類対象に、照応詞が指す先行詞を特定す

るための検出器とを含む。これら検出器は、それぞれ別個の統計的モデルを有する。本実施の形態では、省略検出器 166、照応詞検出器 168、及び先行詞特定器 170 は、いずれも SVM (Support Vector Machine) を統計的モデルとして用いる。

[0026] 照応・省略解析システム 150 はさらに、省略検出器 166、照応詞検出器 168、及び先行詞特定器 170 を用いて、入力テキスト中の省略箇所及び照応詞を検出し、それらの指す先行詞を特定する情報を付して出力する照応・省略解析装置 172 を含む。

[0027] <アノテーションデータ生成補助システム 160>

アノテーションデータ生成補助システム 160 は、表示装置とキーボード及びマウス等を用いてユーザとの間に対話的な入出力を行う入出力装置 202 と、言語知識を含む様々な資源を用いて、テキストアーカイブ 192 に記憶された大量のテキストに対するユーザのアノテーション付与作業を補助することで、学習データを生成し学習データ DB 162 に出力する学習データ生成補助装置 204 とを含む。

[0028] 学習データ生成補助装置 204 が使用する資源は、学習データ DB 162 と同様の学習データを記憶した既存小規模学習データ DB 190 と、テキストアーカイブ 192 に記憶されたテキストに対して適用可能な形式の言語知識を規則の形で記憶した言語知識 DB 194 と、学習データ生成補助装置 204 が言語知識 DB 194 に記憶された言語知識をテキストに対して適用する際に、規則中の表現を他の表現に言換えたり、規則中の単語を別の単語に入れ替えたりするために使用される言換え規則及びカテゴリを記憶した言換え・カテゴリ辞書 196 と、テキストが特定のパターンを持つときに、そのパターンに付随する省略箇所を特定できるよう、予め準備されたパターンを記憶するためのパターン辞書 198 とを含む。

[0029] アノテーションデータ生成補助システム 160 はさらに、学習データ生成補助装置 204 が、テキスト中に省略箇所を発見したときに、その省略箇所が指し示す単語の候補を問い合わせる質問応答システム 206 と、学習デー

タ生成補助装置 204 が質問応答システム 206 に対する質問をテキストに基づいて作成するときに参照する質問タイプ DB 200 と、テキストに対して照応・省略解析を行って、学習データ生成補助装置 204 が実行する照応・省略解析の際に、省略箇所、照応詞、先行詞等の候補を学習データ生成補助装置 204 に対して提示するための既存照応・省略解析器 208 とを含む。

[0030] 既存小規模学習データ DB 190 の構成は、テキスト中の省略箇所、照応詞、及び先行詞の特定ができるようなフォーマットであればどのような形式でもよい。

[0031] テキストアーカイブ 192 に記憶されたテキストは、本実施の形態では単純なテキストであり、形態素解析、構造・係り受け解析がされていないことを前提とする。しかし、本発明はそのような実施の形態には限定されない。他のシステムで形態素解析若しくは構造・係り受け解析又はその双方が行われ、そうした情報が付されたものを用いてもよい。

[0032] 〈言語知識 DB 194〉

図 5 を参照して、言語知識 DB 194 に記憶された言語知識の規則は、大きく分けて 4 種類ある。すなわち、因果関係知識 290、含意関係知識 292、事態生起順序関係知識 294、及び事態生起無効化関係知識 296 である。各知識は、ワイルドカード的な役割を果たす、任意の名詞句等に置換可能な変数（以下では、「X」、「Y」等によって表わす）を含んでもよい。

[0033] 因果関係知識 290 は、例えば「X がタバコを吸う」という、原因に相当する表現と、「X が肺がんの可能性を高める」という、結果に相当する表現との順序付ペアである。ここでは、原因に相当する表現が先に、結果に相当する表現が後に位置するものとする。

[0034] 含意関係知識 292 は、例えば「X が Y の著者だ」という表現により、「X が Y を執筆した」という表現が含意されている、ということをこれら 2 つの表現の順序付ペアで表す。ここでは、含意する表現が先に、含意される表現が後に位置するものとする。

[0035] 事態生起順序関係知識 294 は、例えば「XがYを発見する」という事象が生じると、「XがYを発表する」という事象が生じ得る、という事象の正規の順序関係を、これら2つの表現の順序付ペアで表す。ここでは、事象の順序関係にしたがって表現が配置されるものとする。

[0036] 事態生起無効化関係知識 296 は、例えば「Xを禁止する」という事象が発生すると、「Xを公開できない」という事象が発生するように、ある事象が発生したことにより別のある事象の発生が無効化されるような関係にある表現の順序付ペアである。ここでは、発生する事象が先に位置し、その事象により発生が無効化される事象が後に位置するものとする。

[0037] 後述するように、これら知識は単独でテキストに対して適用できるが、それだけではなく、例えば2つの知識のうち、一方の知識の後半の表現が、他方の知識の前半の表現と一致するような場合には、知識を連鎖させて適用できる。なお、ここでいう「一致」は、例えば同義語の言換え、含意等も含む。

[0038] 〈質問タイプDB200〉

図6を参照して、質問タイプDB200は、学習データ生成補助装置204が質問応答システム206を利用して先行詞の候補を入手するときに利用される。質問タイプDB200は、複数のエントリを持つ。各エントリは例えば、ある動詞について、ある表現内でのその動詞の係り受け関係において省略されている部分の文法役割と、省略部分に挿入されるべき表現を求める際に、どのような形で質問文を生成すればよいかを示す情報（質問タイプ）とを含む。例えば「食べる」という動詞について、ある表現内でその主語が省略されている場合には、「誰が」を文の先頭に配置して文を生成すれば、回答として「食べる」の主語になるべき表現が質問応答システムから得られる。

[0039] 〈質問応答システム206〉

図7を参照して、質問応答システム206を利用して省略箇所の指す表現の候補を得るプロセスについて説明する。まず、入力テキスト320が2つ

の文を含むものとする。第1の文は「夕食に赤ワインを飲みましょう」という文であり、第2の文は「心臓病を防ぎます」であるものとする。第2の文の先頭には「防ぎます」という動詞332の主語が省略されており、この省略箇所334には省略を示すタグが付されているものとする。ここでは、この省略箇所には「赤ワイン」という単語330が入るべきである。

[0040] ここで、このテキストから省略箇所334に入るべき表現の候補を質問応答システムから得るためには、適切な質問文を生成して質問応答システム206への入力として与える必要がある。ここでは、入力テキスト320の省略箇所の文で「防ぎます」の主語を知る必要がある。図示しないが、言換え・カテゴリ辞書196に「予防」と「防ぐ」との言換えが可能であるという情報が格納されているものとする。すると、図6に示す質問タイプDB200に、動詞が「予防する」、文法役割が「主語」となったエントリが見つかる。その「質問タイプ」のカラムには「何が」が格納されている。学習データ生成補助装置204は、これら情報から、「何が心臓病を防ぐのか」という質問文336を生成する。質問文の生成において、文末は質問文にふさわしく、予め準備された規則にしたがって変換する。文末の変換規則を、質問タイプDB200の各エントリに予め格納しておくようにしてもよい。

[0041] 図7を参照して、質問応答システム206は、この質問文336を受け取ると、システム内のデータベース内から、質問文336に対する回答としてふさわしい単語候補群342を検索して取り出す。質問応答システム206はさらに、このようにして得られた単語候補群342を、カテゴリ別に、各カテゴリ344、346、及び348が1又は複数の回答候補を含むようにして学習データ生成補助装置204に送信する。ここでは、カテゴリ344が「赤ワイン」、「酒」、「ビール」を含むものとする。他のカテゴリ346及び348も同様である。学習データ生成補助装置204は、例えば、こうして得られた単語の候補のうち、入力テキスト320内で省略箇所334より前の位置にある単語と一致する表現（この例では「赤ワイン」という単語330）を選択し、省略箇所334と、その指し先の候補としての単語3

30とをアノテーション350の対象とする。以上が質問応答システム206を用いたアノテーション付加作業の概略である。

[0042] 〈既存照応・省略解析器208〉

図8を参照して、既存照応・省略解析器208は、既存の、照応・省略解析を行うものである。ここでは、既存照応・省略解析器208は、省略箇所候補、照応詞候補、及び省略箇所及び照応詞の先行詞候補を提示するために使用される。最終的には人間がこれら候補の中から正しい候補を選択する。したがって、ここで用いる既存照応・省略解析器208の精度は必ずしも高くなくてもよい。

[0043] 〈学習データ生成補助装置204〉

図8を参照して、学習データ生成補助装置204は、テキストアーカイブ192に含まれる各テキストに対して形態素解析を行い、さまざまな文法情報が付された形態素列を出力する形態素解析システム380と、形態素解析システム380の出力する形態素列に対し、文の構造及び係り受け解析を行い、構造・係り受け情報が付された形態素列を出力する係り受け関係解析システム382と、係り受け関係解析システム382が出力する形態素列を、それらの文法情報及び構造・係り受け関係とともに記憶する解析後テキストDB384とを含む。ここでは、係り受け関係解析システム382は、少なくともどの単語がどの単語に係っているかを示す係り受け情報を各形態素列に付すものとする。なお、図8においては、図面を簡略にするために図示していないが、本実施の形態では、学習データ生成補助装置204内の各要素は、入出力装置202を介したユーザの指示により動作する。

[0044] 学習データ生成補助装置204はさらに、解析後テキストDB384に記憶された解析後テキストにおいて、省略箇所を検出し、質問タイプDB200及び質問応答システム206を用いて各省略箇所の先行詞を指すアノテーション候補群を生成する第1の候補生成部388と、解析後テキストDB384に記憶された解析後テキストにおいて、第1の候補生成部388と同様に省略箇所を検出し、各省略箇所に対して言語知識DB194、言換え・カ

テゴリ辞書 196、及びパターン辞書 198 を用いて省略箇所に関するアノテーション候補群を生成する第 2 の候補生成部 390 と、既存小規模学習データ DB 190 に記憶された学習データに対して、その中に存在している照応関係、省略関係の一部を変更して新たな学習データを生成し、その学習データから新たなアノテーション候補を生成する第 3 の候補生成部 392 と、既存照応・省略解析器 208 を用い、解析後テキスト DB 384 に記憶されたテキストに対する照応・省略解析を行って、その結果をアノテーション候補として生成する第 4 の候補生成部 394 とを含む。

[0045] 学習データ生成補助装置 204 はさらに、第 1 の候補生成部 388、第 2 の候補生成部 390、第 3 の候補生成部 392、及び第 4 の候補生成部 394 の出力するアノテーション候補群を、各アノテーションが付されるべき省略箇所、照応詞等と関連付けて記憶する候補 DB 386 と、解析後テキスト DB 384 に記憶された各テキストの省略箇所及び照応詞の各々について、候補 DB 386 を参照して候補を選択可能な形式で入出力装置 202 の表示装置に表示し、ユーザからの選択入力を受け付けて学習データ DB 162 に学習データとして出力する対話型アノテーション装置 396 とを含む。

[0046] 再び図 4 を参照して、学習システム 164 は、アノテーションデータ生成補助システム 160 で用いられた質問タイプ DB 200 及び学習データ生成補助装置 204 とそれぞれ同様の質問タイプ DB 220 及び質問応答システム 226 と、後述するように、先行詞特定器 170 の学習の際に、選択される先行詞に課される制約を記述した選択制限 DB 224 と、アノテーションデータ生成補助システム 160 で使用される言語知識 DB 194 と同様の構成を持つ言語知識を記憶した言語知識 DB 228 と、学習データ DB 162 に記憶された学習データ、質問タイプ DB 220、質問応答システム 226、言語知識 DB 228、及び選択制限 DB 224 を用い、省略検出器 166、照応詞検出器 168、及び先行詞特定器 170 の学習処理を実行する検出器学習装置 222 とを含む。

[0047] 〈第 1 の候補生成部 388〉

図9を参照して、第1の候補生成部388は、解析後テキストDB384に記憶された各テキストデータのうちの述語を検索する述語検索部420と、述語検索部420が出力した各述語のリストを各述語の出現位置とともに記憶する述語リスト記憶部422と、述語リスト記憶部422に記憶された各述語のうち、当該述語が関与する係り受け関係のうちで省略されている箇所を検出し、省略候補として出力する省略候補検出部424と、省略候補検出部424に記憶された省略候補の各々に対して、質問タイプDB200を参照してその省略箇所を補完する単語を回答として求める質問文を生成し、質問応答システム206に質問文として与える質問文自動生成部426と、質問文自動生成部426から質問文を、その質問文に対する回答候補群を質問応答システム206から受信し、質問文とその回答候補群とを組み合わせ出力する回答受信部428と、回答受信部428が出現する質問文と回答候補群とのペアを受け、回答候補群に含まれる候補のうち、解析後テキストDB384に記憶された解析後テキスト文内の出現箇所を確認し、質問の対象となっている省略箇所より前の位置に出現するものを全て選択して省略箇所の指し先候補として出力する出現位置確認部430と、出現位置確認部430から出力される省略箇所とその省略箇所を補完する候補群とを組にしてアノテーション候補として候補DB386に追加する候補追加部432とを含む。

[0048] 〈第2の候補生成部390〉

図10を参照して、第2の候補生成部390は、図9に示す述語検索部420、述語リスト記憶部422、及び省略候補検出部424とそれぞれ同様の述語検索部450、述語リスト記憶部452、及び省略候補検出部454と、省略候補検出部454により検出された省略候補のうち、テキスト内でその省略の候補を持つ述語よりも前に出現している他の述語を検索し、検索された各述語とその省略の候補を持つ述語とをペアにして出力する対象述語検索部456と、対象述語検索部456が出力する各述語ペアについて、それらペアを構成する述語、又はそれを言換え・カテゴリ辞書196を参照し

て得られた等価な表現で言換えた述語が、第1文と第2文とにそれぞれ出現するような言語知識が言語知識DB194内にあるか、又はそうした表現のパターンがパターン辞書198にあるか否かを判定し、あればその言語知識又はパターンを出力する言語知識検索部458と、言語知識検索部458が出力する言語知識又はパターンを一時記憶する合致言語知識記憶部460と、合致言語知識記憶部460に記憶された言語知識又はパターンを用い、対象述語検索部456が出力した述語ペアの各述語を含む表現を用いて、それら表現に含まれる省略箇所の指し先を推定し、省略箇所と指し先とを組にしてアノテーション候補として候補DB386に記憶させる候補追加部462とを含む。

[0049] 候補追加部462による省略箇所の指し先の推定は以下のようにして行われる。図11を参照して、テキスト482中出现するある述語対（文字列494の「導入する」、文字列500の「設置する」）に対して、言語知識検索部458が言語知識480を検索したものとする。テキスト482は、図2に示す例文60と同文である。言語知識480は、「XがYを導入する」という第1文と、「XがYを設置する」という第2文とからなる。テキスト482内で、第1文の変数Xに相当する表現は「N社は」という文字列490である。第1文の「Y」に相当する表現は、テキスト482内の「新型交換機を」という文字列492である。そして、言語知識480の第1文の「導入する」という表現は、テキスト482内の「導入する」という文字列494と一致している。この関係から、変数Xが指す候補は文字列490中の「N社」という単語であり、変数Yが指す候補は文字列492中の「新型交換機」という単語であることが分かる。

[0050] 一方、言語知識480の第2文の「設置する」という表現と合致する表現としてテキスト482中出现するのは、「200システムを設置する」の中の「設置する」という文字列500である。この表現の先頭には省略箇所496を示すタグが付されている。この文（「（〈主語〉は）200システムを設置する」）という文と言語知識480の第2文とを比較することにより

、Yは「システム」に相当し、Xは省略されていることがわかる。言語知識480の第1文から得られた知識を用いると、テキスト482の省略箇所496は、言語知識480のXを指し、したがって「N社は」という文字列490が「設置する」の主語であることが分かる。さらに、テキスト482の第2文の「システム」が言語知識480の変数Yに相当し、したがってテキスト482の「新型交換機は」という文字列492がテキスト482の「システム」という文字列498に対応するものであることが分かる。このようにして、省略箇所496の指し先の候補が言語知識480とテキストとを比較することにより求められる。

[0051] 言語知識を用いた別の例について、図12を参照して説明する。この例では、2つ以上の規則を順次結合することにより、3つ以上の文についての連続した因果関係が得られ、その関係を用いることで、テキスト中に出現する3つ以上の述語に関連して検出される2か所以上の省略箇所の指し先候補を特定する処理を表す。

[0052] 図12に示すように、第1の言語知識520として「砂漠化が進む→黄砂が増える」、第2の言語知識522として「黄砂が増える→喘息が悪化する」という知識があるものとする。第1の言語知識520の第2文と、第2の言語知識522の第1文とは一致する。したがって、これら2つの知識を連鎖させることにより、「砂漠化が進む→黄砂が増える→喘息が悪化する」という第3の知識524が得られる。この第3の知識524とたとえば図12に示すテキスト526とを、特に述語その出現順序に注目して比較する。ここで、テキスト526については省略箇所の推定がされており、省略箇所542、省略箇所548等を示すタグが付加されているものとする。なお、テキスト526において、「(φ1)」は第1の省略箇所を、「(φ2)」は第2の省略箇所を、それぞれ表し、以下同様である。すると、テキスト526に「砂漠化が進んでいて」という表現540、「(φ1が) 増えている」という表現544、及び「(φ2が) 悪化して」という表現546が、第3の知識524の3つの述語部分と同じ順序で出現している。これらの表現の

述語部分は、第3の知識524の述語と同じか、その活用形である。そこで、第3の知識524とテキスト526のうちで上記述語の出現している箇所とを比較することにより、 $\phi 1$ は黄砂を指し、 $\phi 2$ は喘息を指すことが分かる。

[0053] このような比較により、2つ以上の知識を連結して得た知識を活用することで、省略箇所の指し先を効率よく探索できる。なお、知識をどの程度まで連結するかは、設計事項に関する。知識が長くなると1つの知識で探索可能な省略箇所の指し先数の最大値は大きくなるが、規則が適用可能なテキストは少なくなる。本実施の形態では、2つの知識までの連結することを想定している。ただし、本発明はそのような実施の形態には限定されない。もっと多くの知識を連結して省略箇所の指し先の探索に用いてもよい。又は、文の種類により連結数の最大値を変化させたり、ユーザに知識の連結数の最大値を指定させたりしてもよい。

[0054] 図13を参照して、知識の述語の言換えを適用することにより、一つの知識でより多くの指し先の探索を行う例について説明する。知識として「Xが生まれる→Xを販売する」という事態生起順序関係言語知識570があるものとする。一方、「販売する」を「売る」と言換えられるという言換え規則574が言換え・カテゴリ辞書196に記憶されているものとする。処理対象のテキストとして、「ウィスキー『W』が北海道で生まれA社が($\phi 1$)販売した」というテキスト572を考える。処理対象のテキストは、($\phi 1$)で表される省略箇所582を含む。テキスト572の前半部分と事態生起順序関係言語知識570の第1文とを比較すると、変数Xがテキスト572中の「ウィスキー『W』」という単語580を指すことが分かる。一方、テキスト572の述語「販売する」の目的語がXとなっている。この結果と、テキスト572の後半部の述語が「販売した」という表現584であることから、テキスト572中の「 $\phi 1$ 」で示される省略箇所582が「ウィスキー『W』」という単語580に相当することが分かる。そこで、まずこのようなアノテーション済のテキスト572が得られる。さらに、こうして得ら

れたテキスト 5 7 2 に対し、言換え規則 5 7 4 を適用した言換え 5 7 6 により「販売した」を「売った」という表現 5 8 6 に置換し、アノテーションはそのまま残すことで、アノテーション済の新たなアノテーション済テキスト 5 7 8 が得られる。

[0055] 図 1 4 を参照して、さらに、経験的に得られた文のパターンを用いて省略箇所又は照応詞の指し元及び／又は指し先を特定する処理について説明する。例えば、「Xは～して、～する」のようなパターンが出現する場合、「〈述語〉して」の主語と「〈述語〉する」の主語とがいずれも X であるような例が多いことが経験的に知られている。そうした知識を文法的なパターンとしてパターン辞書 1 9 8 に記憶しておく。パターン中には、省略箇所に相当する部分を特定する情報と、その指し先を示す情報とを挿入しておいてもよい。そして、処理対象のテキスト中にパターン辞書 1 9 8 に記憶されたパターンに合致するような表現が出現した場合、その表現をパターンと比較することで、入力箇所内の省略箇所（指し元）と、その省略箇所の指し先との候補を特定できる。

[0056] 例えば、図 1 4 に示すように、パターン 6 1 2 がパターン辞書 1 9 8 に記憶されているものとする。このパターン 6 1 2 は、「Xは」という表現 6 5 0、「…して」という表現 6 5 2、及び「…する」という表現 6 5 4 がこの順序で出現した場合、表現 6 5 4 の先頭に表現 6 5 4 の主語が省略された箇所を示すタグ 6 5 6 が付されること、そのタグ 6 5 6 の指先が X であることが記述されているものとする。

[0057] 入力としてテキスト 6 1 0 が与えられたものとする。テキスト 6 1 0 には、「台風 2 0 号は」という表現 6 3 0、「…を北上し」という表現 6 3 2、及び「接近する」という表現 6 3 4 がこの順番で出現する。このテキスト 6 1 0 をパターン 6 1 2 と照合 6 1 4 することにより、X が台風 2 0 号に相当すること、「接近する」の主語が省略されており、その指し先は台風 2 0 号であることが判明する。そこで、「接近する」の直前の省略箇所を示すタグ 6 3 6 と、その指し先としての「台風 2 0 号」とを対にしてアノテーション

候補とする。このように、人の経験から導き出されたパターンを照応・省略解析に適応することで、人が持つ知識を生かした照応・省略解析を実現できる。

[0058] 〈第3の候補生成部392〉

図15を参照して、第3の候補生成部392は、既存小規模学習データDB190の中からアノテーション済の照応関係及び省略関係を検索し、両者を区別して出力する照応・省略関係検索部680と、照応・省略関係検索部680の出力する照応関係を構成する名詞句対の各々について、その名詞句の内の指し元（テキスト内で後方に位置するもの）を自動的に省略した新たな文を生成して新たなテキストを生成する指し元省略文生成部682と、指し元省略文生成部682の生成した文のうち、省略された箇所と、省略された箇所とともに照応関係がアノテーションされていた名詞句との間に新たな省略関係のアノテーション候補を付すアノテーション追加部684と、照応・省略関係検索部680が出力する省略関係を含むテキスト部分のうち、省略箇所について、アノテーションに基づいて省略箇所に、その指し先の要素を代名詞として復元することで新たな文を生成する省略要素復元部688と、省略要素復元部688により生成された新たな文に、省略要素復元部688により復元された代名詞の指し元とその指し先とからなるアノテーション候補を追加するアノテーション追加部690と、アノテーション追加部684及びアノテーション追加部690により生成された新たな文をそのアノテーション候補とともに候補DB386に追加する候補追加部686とを含む。

[0059] 第3の候補生成部392により指し元を省略した場合、省略は機械によりある基準にしたがって自動的に行われる。そのため、省略語の表現について、その省略部分の指し先が何かを人が判断できない場合も生ずる。そうした場合には、そのアノテーションを負例として学習データに追加する。

[0060] なお、候補追加部686は、照応・省略関係検索部680及び省略要素復元部688の出力するテキストが解析後テキストDB384に存在しない場

合には、これらテキストを解析後テキストDB384に追加し、アノテーション候補を解析後テキストDB384に追加されたテキストと関連付けて候補DB386に追加する。

[0061] 〈第4の候補生成部394〉

図16を参照して、第4の候補生成部394は、解析後テキストDB384に記憶された解析後テキストに対して、既存照応・省略解析器208を適用することにより、照応・省略解析の実行結果を得て、解析後テキストにアノテーションとして追加する照応・省略解析実行部710と、照応・省略解析実行部710の出力する照応・省略解析結果のアノテーションが付されたテキストを記憶する解析結果記憶部712と、解析結果記憶部712に記憶された解析結果に含まれるアノテーションを、解析後テキストDB384で解析対象となったテキストに対するアノテーション候補として候補DB386に追加する候補追加部714とを含む。

[0062] 〈対話型アノテーション装置396〉

図8に示す対話型アノテーション装置396は、図17に示すような状態遷移を行うプログラムにより実現される。図17を参照して、このプログラムは、プログラム実行開始時及び文書クローズ時等に初期処理を実行し、初期画面を表示した状態（初期状態）740から開始する。この状態では、テキストアーカイブから処理対象の文書を選択する処理と、プログラムの実行を終了する処理とのいずれかを選択できる。文書の選択が選ばれるとプログラムの状態は文書選択状態742に遷移する。文書選択状態742では、文書ファイルの選択ダイアログが表示される。ここでは文書ファイルの選択と、処理のキャンセルとを選ぶことができる。処理のキャンセルが選ばれると状態は初期状態740に戻る。文書選択状態742で文書が選択されるとプログラムは文書の内容を表示する状態（文書表示状態）744に遷移する。文書表示状態744では、文書の表示をキャンセルし、更新を反映せずに再度文書選択状態742に戻る処理と、更新を反映して文書をクローズする処理と、照応・省略解析の学習データを作成するために、アノテーション候補

の選択を指示する処理とのいずれかを選択できる。アノテーション候補の選択が指示されると、文書の、現在処理中の位置から文書の末尾方向に向かって、アノテーション候補を検索する状態（候補検索状態）746を経由した後、アノテーション候補を表示してユーザによるアノテーション候補の選択を待つ状態（候補選択待機状態）748に遷移する。

[0063] 候補検索状態746では、アノテーション候補が付された次の位置を検索し、検索された位置に付されているアノテーション候補が1つだけであれば図3の下段に示したような、アノテーション候補を承認するか否かを指定するドロップダウンリストを生成し、ドロップダウンリスト表示のためのアイコン「▼」を対象箇所に表示する。アノテーション候補が複数個ある場合には、それらを全て表示するドロップダウンリストを生成し、ドロップダウンリスト表示のためのアイコンを対象箇所に表示する。なお、ドロップダウンリストの表示時に、ユーザがリストのいずれかにポインタを合わせると、その候補と、その候補に対応する文書上の表現とが同じ色でハイライトされる。こうした処理により、指し元と指し先との対応についてユーザが容易に理解できる。この状態では、アノテーション選択処理を中断して文書表示状態744に状態を遷移させる処理と、それまでに行われたアノテーションの選択を文書に反映させて文書を保存し、文書ファイルをクローズする処理とのいずれかが選択できる。

[0064] ドロップダウンリストの操作により候補のいずれかが選択がされると、プログラムはアノテート付加状態750に遷移する。アノテート付加状態750では、選択にしたがって文書内の指定箇所を選択されたアノテートを付し、他の候補には選択されなかったことを示す情報を付すよう、メモリ上の文書を更新する。プログラムは再び候補検索状態746に移動し、次の候補が見つかり候補選択待機状態748に遷移する。

[0065] 候補選択待機状態748で文書を表示させたままアノテーション処理を中断することを選択すると、プログラムは文書表示状態744に遷移する。文書に対して行ったアノテーション処理を反映させて現在の文書を保存するこ

とを選択した場合、プログラムは、文書クローズ状態 7 5 2 に遷移する。文書クローズ状態 7 5 2 では、プログラムは、メモリにロードされている更新済の文書データを文書ファイルとして上書き保存し、開いている文書ファイルをクローズする処理と、文書を保存せず、文書選択状態 7 4 2 に遷移する処理とのいずれかを選択できる。文書を保存しないことが選択されると、状態は文書選択状態 7 4 2 に直接遷移する。文書を保存することが選択されると、メモリ上のファイルを文書ファイルとして記憶媒体に上書き保存する状態（上書保存状態） 7 5 4 に遷移し、保存が完了すると文書選択状態 7 4 2 に遷移する。

[0066] このプログラムを実行することにより、省略箇所、照応詞を装置側で自動的に検索し、いずれかが検索されるたびに、その箇所に付されたアノテーション候補が表示される。アノテーション候補が 1 つであればユーザはそのアノテーション候補を承認するか否かを入力すればよい。場合によってはアノテーション候補を手入力してもよい。いずれにせよ、削除位置又は照応詞をユーザが目視で探し、その指し先をさらに目視で探す場合と比較して、はるかに容易にアノテーションを行える。

[0067] <検出器学習装置 2 2 2>

再び図 4 を参照して、検出器学習装置 2 2 2 は、前述したとおり、学習データ DB 1 6 2 に記憶された学習データから、選択制限 DB 2 2 4、質問タイプ DB 2 2 0 及び質問応答システム 2 2 6 を用いて省略検出器 1 6 6、照応詞検出器 1 6 8、及び先行詞特定器 1 7 0 の学習をそれぞれ別々に行う。先行詞特定器 1 7 0 についてはさらに、省略箇所の先行詞用の特定器と、照応詞の先行詞用の特定器との学習を独立に行う。

[0068] 図 1 8 を参照して、検出器学習装置 2 2 2 は、このために、省略検出器学習部 7 7 0、照応詞検出器学習部 7 7 2、及び先行詞特定器学習部 7 7 4 を含む。

[0069] <省略検出器学習部 7 7 0>

本実施の形態では、省略検出器 1 6 6 は、主語省略検出器 8 0 0、目的語

省略検出器 802, 及び間接目的語省略検出器 804 を含む。省略検出器学習部 770 は、これら 3 つの検出器の学習を個別に行うため、学習データ DB 162 に記憶された学習データの省略箇所の各々が、述語の主語（ガ格）、直接目的語（ヲ格）、又は間接目的語（二格）であるときに、それらを分類して、主語省略検出器 800、目的語省略検出器 802、及び間接目的語省略検出器 804 の学習をそれぞれ行うための素性ベクトル群 782、784 及び 786 を生成するための格別素性ベクトル生成部 780 と、これら素性ベクトル群 782、784 及び 786 を用いてそれぞれ主語省略検出器 800、目的語省略検出器 802、及び間接目的語省略検出器 804 を構成する SVM の学習を行うための SVM 学習処理部 788 とを含む。

[0070] ここでは、述語とは、動詞、形容詞、及び「名詞＋『だ』」の形式の文字列である。素性ベクトルに含む学習のラベルとしては、述語が省略を持つ場合に 1、それ以外なら 0 を用いる。

[0071] 格別素性ベクトル生成部 780 が生成する素性ベクトルは、以下のような要素を含む。

- ・省略検出対象の述語に関して、検出対象の文法役割（例えば、主語）が係り受け関係にあるなら 1、それ以外なら 0
- ・主題となる表現（「は」でマークされた名詞句）が述語と同一文内に出現しているなら 1、それ以外なら 0
- ・省略検出対象の述語が対象テキストの 1 文目に出現しているなら 1、それ以外なら 0
- ・省略検出対象の述語が文の最初の後であるなら 1、それ以外なら 0
- ・省略検出対象の述語と係り受け関係にある語の見出し語、品詞

[0072] 〈照応詞検出器学習部 772〉

照応詞検出器学習部 772 は、学習データ DB 162 に記憶された学習データから、照応詞検出器 168 の SVM の学習を行うために、照応関係の差し元（照応詞）としてアノテーションが付された単語を選択する照応詞候補選択部 830 と、照応詞候補選択部 830 により選択された照応詞の各々に

ついて、後述するような素性ベクトルを生成するための素性ベクトル生成部 832 と、素性ベクトル生成部 832 により生成された素性ベクトル群を用い、照応詞検出器 168 を構成する SVM の学習を行うための SVM 学習処理部 834 とを含む。

[0073] 素性ベクトル生成部 832 が生成する素性ベクトルにおいて、学習のラベルは、分類対象の名詞句が、テキストの前方に照応関係となる差し先を持つなら 1、それ以外なら 0 である。照応詞検出学習のための素性ベクトルの要素は以下を含む。

- ・照応詞候補の品詞、見出し語文字列、及びその候補に後続する格助詞
- ・照応詞候補の係り先の品詞、見出し語、その係り先に後続する格助詞
- ・テキスト中で照応詞候補より前の位置に出現する名詞句が文字列として照応詞候補と完全一致する場合に 1、それ以外は 0
- ・テキスト中で照応詞候補と部分一致する場合に 1、それ以外は 0

[0074] 〈先行詞特定器学習部 774〉

先行詞特定器 170 は、照応詞用先行詞特定器 900 と、省略用先行詞特定器 902 とを含む。そのため、先行詞特定器学習部 774 はこの 2 つの特定器 900 及び 902 の学習を別個に行う構成を有する。

[0075] 具体的には、先行詞特定器学習部 774 は、照応詞用先行詞特定器 900 の学習を行う照応詞用先行詞特定器学習部 840 と、省略用先行詞特定器 902 の学習を行う省略用先行詞特定器学習部 842 とを含む。

[0076] 照応詞用先行詞特定器学習部 840 は、学習データ DB 162 に記憶された学習データから照応関係の指し元（照応詞）としてアノテーションが付された表現を選択する照応詞選択部 850 と、照応詞選択部 850 が選択した照応詞に対し、実際に先行詞としてアノテーションされたものとそれ以外の先行詞の可能性のある表現とからなる先行詞候補を選択する先行詞候補選択部 852 と、照応詞選択部 850 が選択した照応詞と、先行詞候補選択部 852 が選択した照応詞候補との組み合わせの各々について照応詞用先行詞特定器 900 の学習を行うための素性ベクトルを生成する素性ベクトル生成部

８５４と、素性ベクトル生成部８５４が生成した素性ベクトルを用いて照応詞用先行詞特定器９００を構成するＳＶＭの学習を行うためのＳＶＭ学習処理部８５６とを含む。

[0077] 一方、省略用先行詞特定器学習部８４２は、学習データＤＢ１６２内の学習データのうち、省略関係としてアノテーションされた表現（省略）を選択する省略選択部８７０と、省略選択部８７０が選択した省略に対する先行詞としてアノテーションされた表現を含め、先行詞となる可能性のある候補を学習データ中で選択する先行詞候補選択部８７２と、省略選択部８７０が選択した省略部と、先行詞候補選択部８７２が選択した先行詞候補との組み合わせの各々について、素性ベクトルを生成するための素性ベクトル生成部８７４と、素性ベクトル生成部８７４により生成された素性ベクトルを用いて、省略用先行詞特定器９０２を構成するＳＶＭの学習を行うＳＶＭ学習処理部８７６とを含む。なお、素性ベクトル生成部８７４は後述するように選択制限ＤＢ２２４、質問タイプＤＢ２２０、質問応答システム２２６、及び言語知識ＤＢ２２８に接続されており、素性ベクトルの生成時にこれらから得られる情報を用いる。

[0078] 素性ベクトル生成部８５４及び素性ベクトル生成部８７４が生成する素性ベクトルにおいて、学習のラベルは、分類対象の省略・先行詞候補又は照応詞・先行詞候補の対が照応・省略関係となるなら１，それ以外は０である。

[0079] 素性ベクトルの要素は以下のとおりである。

- ・先行詞候補の品詞、見出し語文字列、その候補に後続する格助詞
- ・先行詞候補がテキストの１文目に出現するなら１，それ以外なら０
- ・先行詞候補が文内で最初に言及された先行詞候補なら１，それ以外なら０
- ・先行詞候補と照応詞が完全に同じ文字列なら１，それ以外なら０
- ・先行詞候補と照応詞が部分的に同じ文字列なら１，それ以外なら０
- ・選択制限ＤＢ２２４を用い、省略を持つ述語と先行詞となる名詞句の間で選択制限が満たされる場合には１，それ以外なら０

・省略を含む文、その省略の箇所について、質問タイプDB220を用いてその箇所が回答となるような質問文を生成し質問応答システム226に与えて得られた回答のいずれかが先行詞候補と一致すれば1、それ以外なら0

・省略を持つ述語、先行詞の係り先の述語が検出器学習装置222に登録されている因果関係知識に含まれる述語対と合致するなら1、それ以外なら0

[0080] <照応・省略解析装置172>

図4を参照して、照応・省略解析装置172は、省略検出器166、照応詞検出器168及び先行詞特定器170に接続され、これらを用いて入力250に対する照応・省略解析をし、自動的にアノテーションを入力250に追加して出力254とするための照応・省略解析器252と、照応・省略解析器252が照応・省略解析をする際の素性ベクトル生成に使用する言語知識DB256、質問タイプDB258、質問応答システム260、及び選択制限DB262を含む。言語知識DB256は、照応・省略解析システム150で使用される言語知識DB194、学習システム164で使用される言語知識DB228と同様の言語知識を記憶したデータベースである。質問タイプDB258は、アノテーションデータ生成補助システム160で使用される質問タイプDB200及び学習システム164で使用される質問タイプDB220と同様の構成を持ち、その記憶内容も同様である。質問応答システム260は、アノテーションデータ生成補助システム160で使用される質問応答システム206及び学習システム164で使用される質問応答システム226と同様の機能を持つシステムである。選択制限DB262は、学習システム164における素性ベクトルの生成時に使用される選択制限DB224と同様のものである。なお、ここで同様の構成を持つと記載したDBは、互いに同一のものでもよいし、互いに異なるものでもよい。

[0081] [動作]

以上に構成を述べた照応・省略解析システム150は以下のように動作する。学習データの作成、検出器の学習、及び入力に対する自動アノテーショ

ンの３つのフェーズに分けて説明する。

[0082] <学習データの作成>

学習データの作成はアノテーションデータ生成補助システム１６０により行われる。この処理に先立ち、既存小規模学習データＤＢ１９０、テキストアーカイブ１９２、言語知識ＤＢ１９４、言換え・カテゴリ辞書１９６、パターン辞書１９８、既存照応・省略解析器２０８、及び質問タイプＤＢ２００を準備しておく必要がある。また、質問応答システム２０６については、予めアノテーションデータ生成補助システム１６０内に準備し学習データ生成補助装置２０４からアクセス可能にしておくか、外部の質問応答システムに対して質問文を送れるよう準備しておく必要がある。

[0083] 図８を参照して、利用者は学習データ生成補助装置２０４を起動し、テキストアーカイブ１９２を指定して学習データの生成の開始を指示する。形態素解析システム３８０は、テキストアーカイブ１９２に記憶された各テキストを読み出し、形態素解析をして種々の文法的情報が付された形態素列を係り受け関係解析システム３８２に与える。係り受け関係解析システム３８２は、与えられた形態素列からなる各文について文法的構造解析及び係り受け解析を行い、構造情報と係り受け情報とが付された形態素列を解析後テキストＤＢ３８４に出力する。こうして、テキストアーカイブ１９２に記憶された各テキストが解析され、解析後テキストが解析後テキストＤＢ３８４に蓄積される。

[0084] 図９を参照して、第１の候補生成部３８８の述語検索部４２０は、解析後テキストＤＢ３８４に記憶された解析後テキストの各々を読み出し、述語を検索して述語リスト記憶部４２２に書き出す。省略候補検出部４２４は、述語リスト記憶部４２２に蓄積された述語リスト内の各述語について、その述語を含む係り受け関係において、その述語の主語、目的語、間接目的語等のいずれかが存在している否かを判定し、存在していないときにはその部分を省略候補として検出し質問文自動生成部４２６に与える。質問文自動生成部４２６は、省略候補の部分が回答となるような質問文から質問文を生成する。

この際、質問文自動生成部426は、処理対象となっている述語と、省略候補の文法役割（主語、目的語、等）とをキーとして質問タイプDB200をアクセスし、質問タイプを読み出す。質問文自動生成部426はさらに、読み出した質問タイプの表現を用いて、省略箇所を含む文を変形して図7に示す質問文336のような質問文を生成する。質問文自動生成部426はこの質問文を質問応答システム206と回答受信部428に与える。回答受信部428は、質問文自動生成部426から質問文を受けると、質問文に対する回答が質問応答システム206から与えられるまで待機する。

[0085] 質問応答システム206は、本実施の形態では、与えられた質問文に対し、いくつかのカテゴリ別に、各カテゴリについて複数の回答候補を生成し、回答受信部428に与える。回答受信部428は、質問応答システム206から回答を受信すると、その回答が質問文自動生成部426から送信されたどの質問文に対するものかを確認し、出現位置確認部430に処理対象となっている述語を特定する情報と回答とを与える。

[0086] 出現位置確認部430は、回答受信部428からの情報に基づき、処理対象の解析後テキストのうち、処理の対象となっている述語より前に出現している省略の指し先候補の位置を確認し、それらを省略箇所の補完候補として、処理対象となっている述語と、候補の各々とを組みにして候補追加部432に出力する。候補追加部432は、出現位置確認部430から与えられた候補を候補DB386に追加する。

[0087] 図10を参照して、第2の候補生成部390は以下のように動作する。述語検索部450は、解析後テキストDB384に格納された解析後テキストの内の各述語を検索し、述語リスト記憶部452に蓄積する。省略候補検出部454は、図8に示す省略候補検出部424と同様、述語リスト記憶部452に蓄積された述語リスト内の各述語について、その述語を含む係り受け関係において、その述語の主語、目的語、間接目的語等のいずれかが存在している否かを判定し、存在していないときにはその部分を省略候補として検出する。省略候補検出部454は、検出された省略候補の位置を示す情報を

対象述語検索部 4 5 6 に与える。対象述語検索部 4 5 6 は、省略候補検出部 4 5 4 により検出された省略候補のうち、テキスト内でその省略の候補を持つ述語よりも前に出現している他の述語を検索し、検索された各述語とその省略の候補を持つ述語とをペアにして言語知識検索部 4 5 8 に与える。

[0088] 言語知識検索部 4 5 8 は、対象述語検索部 4 5 6 が出力する各述語ペアについて、それらペアを構成する述語又はそれを言換えた述語が言語知識 DB 1 9 4 内の言語知識のいずれかにおいて、第 1 文と第 2 文とに出現するような言語知識があるか否かを判定し、あればその言語知識を出力する。言語知識検索部 4 5 8 はまた、対象述語検索部 4 5 6 から与えられる述語ペアを含む表現であって、かつパターン辞書 1 9 8 に記憶されたパターンに合致するものがあればそれも出力する。言語知識検索部 4 5 8 はこれら言語知識及びパターンをテキストに適用する際に、言換え・カテゴリ辞書 1 9 6 に記憶された言換え規則及びカテゴリを用い、規則中の表現を他の表現に言換えたり、規則中の単語を別の単語に入れ替えたりすることにより、言語知識の適用対象を広げる。言語知識検索部 4 5 8 により検索された言語知識は、その言語知識の検索の際に用いられた述語ペアとともに合致言語知識記憶部 4 6 0 に蓄積される。最後に、候補追加部 4 6 2 が、図 1 1 ～図 1 4 を参照して説明したように、合致言語知識記憶部 4 6 0 に記憶された述語ペアと、合致した言語知識又はパターンとを照合することにより、省略候補の指し先候補を特定し、省略候補と、各指し先候補とのペアをアノテーション候補として候補 DB 3 8 6 に追加する。

[0089] 図 1 5 を参照して、第 3 の候補生成部 3 9 2 の照応・省略関係検索部 6 8 0 は、既存小規模学習データ DB 1 9 0 の中でアノテーション済の照応関係及び省略関係を検索し、照応関係を含む文を指し元省略文生成部 6 8 2 に、省略関係を含む文を省略要素復元部 6 8 8 に、それぞれ出力する。

[0090] 指し元省略文生成部 6 8 2 は、照応関係を含むテキストから、照応関係を構成する名詞句対を探し、その各々について、その名詞句の内、テキスト内で後方に位置する指し元を自動的に省略した新たなテキストを生成してアノ

テーション追加部 684 に与える。アノテーション追加部 684 は、このようにして指し元省略文生成部 682 の生成した文のうち、省略された箇所と、省略された箇所とともに照応関係がアノテーションされていた名詞句との間に新たな省略関係のアノテーション候補を付して候補追加部 686 に出力する。

[0091] 候補追加部 686 は、照応・省略関係検索部 680 が出力する省略関係を含むテキストのうち、省略箇所について、アノテーションに基づいて、その指し先の要素を代名詞として復元することで新たなテキストを生成しアノテーション追加部 690 に与える。アノテーション追加部 690 は、省略要素復元部 688 により生成された新たなテキストに、省略要素復元部 688 により復元された代名詞からなる指し元と、もともとのアノテーションにより指し先とされていた部分とからなる新たな照応関係のアノテーション候補を追加し候補追加部 686 に出力する。

[0092] 候補追加部 686 は、アノテーション追加部 684 及びアノテーション追加部 690 から出力されたテキストをそれらに付加されたアノテーション候補とともに候補 DB 386 に追加する。

[0093] 図 16 を参照して、第 4 の候補生成部 394 の照応・省略解析実行部 710 は、解析後テキスト DB 384 からテキストを読み出し、既存照応・省略解析器 208 による照応・省略解析を実行する。既存照応・省略解析器 208 は、その性能には限界があるものの、入力されたテキストに対して照応・省略解析を実行し、アノテーションされたテキストを照応・省略解析実行部 710 に返す。照応・省略解析実行部 710 は、そのアノテーションされた解析結果を解析結果記憶部 712 に蓄積する。候補追加部 714 は、解析結果記憶部 712 に記憶された解析結果に含まれるアノテーションをアノテーション候補として候補 DB 386 に追加する。

[0094] 以上の処理をテキストアーカイブ 192（図 4 及び図 8 参照）に記憶された全てのテキストに対して実行することにより、最終的に候補 DB 386 にはアノテーション候補が付されたテキストが蓄積される。候補 DB 386 に

蓄積されたテキストは、省略候補及び照応詞候補を含み、各省略候補及び照応詞候補には、それらの指し先を示すアノテーション候補が1又は複数個付加されている。

[0095] 対話型アノテーション装置396は、候補DB386に蓄積された候補の各々について、入出力装置202を用いたユーザとの対話型処理により、省略候補及び照応詞候補の各々について、以下のように動作する。

[0096] 図3の下段を参照して、対話型アノテーション装置396は、アノテーション候補を含むテキスト92を画面に表示する。その際、対話型アノテーション装置396は、省略候補及び照応詞候補の各々について、以下のようにしてアノテーション候補を選択するドロップダウンメニューを作成する。すなわち、省略候補の指し先が1箇所しか存在しない場合には、図3において文字列130で示される省略候補の位置に、その指し先候補である単語110と同じ単語を表示し、さらにその右側にドロップダウンメニューがあることを示すアイコン「▼」を表示する。ドロップダウンメニューには、表示されている候補を受け入れるか否かを示すオプション（YES・NO）と、直接入力とを示すオプションとが表示される。ユーザがYESを選択すると、選択されたアノテーションがアノテーションとして確定される。NOが選択された場合、アノテーションは未確定で残される。直接入力を選択されると、文字列130で示される省略候補の指し先を直接指定するためのダイアログが表示される。ユーザが指し先を入力することにより、ユーザの指定にしたがった新たなアノテーションが生成され、確定アノテーションとなる。この場合、最初に表示されていながらユーザにより選択されなかったアノテーションは、学習時の素性ベクトル作成の際の負例を生成するために使用される。したがって、負例であることを示すフラグがこのアノテーション候補に付される。

[0097] アノテーション候補が複数個ある場合には、それら複数個のアノテーション候補をオプションとして含むドロップダウンリストが生成される。該当箇所には、アノテーション候補のうち先頭のものの、又は何らかの形でアノテ

ション候補にスコア付けした際の最高スコアのアノテーション候補が表示される。その右側には、ドロップダウンリストがあることを示すアイコン「▼」が表示される。この場合にも、ドロップダウンリストには、直接入力することを選択するオプション項目が表示される。ユーザがいずれかのオプションを選択すると、選択されたアノテーション候補がアノテーションとして確定する。選択されなかった候補については、学習時の素性ベクトル作成の際に、負例を生成するために使用される。

[0098] いずれの場合も、アノテーションが確定した解析後テキストは学習データDB162に蓄積される。候補DB386に記憶された全ての解析後テキストについて、上記したアノテーションの選択が完了すると、学習データDB162が完成する。

[0099] <検出器の学習>

図18を参照して、検出器学習装置222は以下のように動作する。省略検出器学習部770の格別素性ベクトル生成部780は、学習データDB162に記憶されたアノテーション確定済の学習データのうち、省略を含む学習データを読み出す。格別素性ベクトル生成部780は、学習データの省略箇所の各々が、述語の主語（ガ格）、直接目的語（ヲ格）、又は間接目的語（ニ格）であるときに、それらを分類して、それぞれ素性ベクトル群782、784及び786を生成し図示しない記憶装置に蓄積する。この際、格別素性ベクトル生成部780は、確定したアノテーションについては学習ラベルを1、それ以外のアノテーションについては学習ラベルを0とする。SVM学習処理部788は、これら素性ベクトル群782、784及び786を用いて、主語省略検出器800、目的語省略検出器802、及び間接目的語省略検出器804が有するSVMの学習を行う。

[0100] 照応詞検出器学習部772の照応詞候補選択部830は、学習データDB162に蓄積された学習データのうち、照応詞候補を含むテキストを読み出し、照応関係の差し元（照応詞）としてアノテーションが付された単語を選択する。素性ベクトル生成部832は、照応詞候補選択部830により選択

された照応詞候補の各々について、前述した素性ベクトルを生成する。この際、素性ベクトル生成部832は、確定したアノテーションについては学習ラベルを1、それ以外のアノテーションについては学習ラベルを0とする。素性ベクトル生成部832が生成した素性ベクトルは図示しない記憶装置に蓄積される。SVM学習処理部834は、このようにして蓄積された素性ベクトル群を用い、照応詞検出器168を構成するSVMの学習を行う。

[0101] 先行詞特定器学習部774において、照応詞用先行詞特定器学習部840は以下のように動作する。照応詞用先行詞特定器学習部840の照応詞選択部850は、学習データDB162に記憶された学習データのうち、照応関係の指し元（照応詞）としてアノテーションされている表現を選択し、先行詞候補選択部852に与える。先行詞候補選択部852は、照応詞選択部850が選択した照応詞に対し、実際に先行詞としてアノテーションされたものとそれ以外の先行詞の可能性のある表現（名詞句）との双方を含む先行詞候補を選択する。素性ベクトル生成部854は、照応詞選択部850が選択した照応詞と、先行詞候補選択部852が選択した照応詞候補の各々との組み合わせについて照応詞用先行詞特定器900の学習を行うための素性ベクトルを生成する。この際、素性ベクトル生成部854は、照応関係において照応詞の指し先としてアノテーションされた表現については学習ラベルを1、それ以外の学習ラベルを0とする。SVM学習処理部856は、素性ベクトル生成部854が生成した素性ベクトルを用いて照応詞用先行詞特定器900を構成するSVMの学習を行う。

[0102] 一方、省略用先行詞特定器学習部842は以下のように動作する。省略用先行詞特定器学習部842の省略選択部870は、学習データDB162内の学習データのうち、省略関係としてアノテーションされた表現（省略）を選択し、先行詞候補選択部872に与える。先行詞候補選択部872は、省略選択部870が選択した省略に対する先行詞としてアノテーションされた表現を含め、省略の先行詞となる可能性のある候補を学習データ中で選択する。素性ベクトル生成部874は、省略選択部870が選択した省略部と、

先行詞候補選択部 872 が選択した先行詞候補の各々との組み合わせについて、素性ベクトルを生成し図示しない記憶装置に蓄積する。このとき、素性ベクトル生成部 874 が生成する素性ベクトルにおいて、学習のラベルは、分類対象の省略・先行詞候補の対が省略関係となるなら 1、それ以外は 0 である。また素性ベクトル生成部 874 は、素性ベクトルの生成にあたって、前述したとおり、選択制限 DB 224、質問タイプ DB 220、質問応答システム 226 及び言語知識 DB 228 を用いる。SVM 学習処理部 876 は、素性ベクトル生成部 874 により生成された素性ベクトルを用いて、省略用先行詞特定器 902 を構成する SVM の学習を行う。

[0103] 以上の処理により、省略検出器 166 に含まれる主語省略検出器 800、目的語省略検出器 802、及び間接目的語省略検出器 804、照応詞検出器 168、並びに先行詞特定器 170 に含まれる照応詞用先行詞特定器 900 及び省略用先行詞特定器 902 の学習が完了する。

[0104] <自動アノテーション>

図 4 に戻り、照応・省略解析器 252 による入力 250 に対する照応・省略解析とその結果の自動アノテーションは以下のようにして実行される。なお、この解析は省略解析、照応解析、省略及び照応詞に対する先行詞解析に分割される。省略解析では省略検出器 166 が用いられる。照応解析では照応詞検出器 168 が用いられる。先行詞解析では先行詞特定器 170 が用いられる。

[0105] 最初に、入力 250 に対する形態素解析及び構造・係り受け解析が実行され、構造・係り受け情報が付された形態素列が処理の対象となる。この形態素列に対し、省略解析では、それぞれ図 18 に示す主語省略検出器 800、目的語省略検出器 802、及び間接目的語省略検出器 804 の学習を行ったときと同様の構成の素性ベクトルを生成し、これら検出器に与える。主語省略検出器 800、目的語省略検出器 802、及び間接目的語省略検出器 804 の出力により、入力 250 のうちでそれぞれ最も高いスコアを示す部分に省略箇所としてのアノテーションが付される。照応詞解析では、入力 250

に含まれる名詞句対中から照応詞候補となり得る語句を選択し、その各々について、図 18 に示す素性ベクトル生成部 832 により生成された素性ベクトルと同様の構成の素性ベクトルを入力 250 から生成する。この組成ベクトルを照応詞検出器 168 に与えることにより、照応詞検出器 168 からスコアが得られる。このスコアにより、照応詞か否かを決定し、その旨のアノテーションを付する。

[0106] 照応詞用先行詞特定器 900 による先行詞の推定は以下のようにして行われる。照応詞検出器 168 により照応詞と判定された表現の各々について、先行詞となり得る表現を入力 250 中で選択し、両者をペアとして図 18 の素性ベクトル生成部 854 と同様に素性ベクトルを生成する。この素性ベクトルを照応詞用先行詞特定器 900 への入力とし、照応詞用先行詞特定器 900 のスコアを得て、しきい値より高いスコアで、かつ最も高いスコアを示した先行詞候補と、処理対象の照応詞との間で照応関係のアノテーションを追加する。

[0107] 省略用先行詞特定器 902 による先行詞の推定も同様である。ただしこの場合、素性ベクトル生成部 854 により生成される素性ベクトルではなく素性ベクトル生成部 874 により作成されるものと同様の構成の素性ベクトルが使用される。したがってこのときには、選択制限 DB 224、質問タイプ DB 220、質問応答システム 226 及び言語知識 DB 228 が必要となる。

[0108] 以上のように本実施の形態によれば、テキストアーカイブから、照応・省略解析のための検出器の学習を行うための学習データ生成が容易に行える。例えば照応・省略関係の候補が自動的に検索され、その先行詞候補がリストとして表示される。ユーザは、リストに表示された候補が 1 つでかつ正しいものであればそれを承認し、リストに表示された候補が複数個であればそのうちの 1 つを選択すればよい。従来のように、照応・省略関係の位置を目視で探し、次にその指し示す位置をさらに目視で探して両者に照応・省略関係のアノテーションを付す、という作業を行う必要はない。仮に表示されたり

ストに正しい指し先が示されないときには、手入力により正しい指し先を指定することができる。

[0109] したがって、従来の技術よりもはるかに簡便に、照応・省略解析の検出器の学習を行うための学習データを準備できる。その結果、学習データの作成コスト及び作成時間をいずれも低減できる。大量の学習データを低コストで準備できるため、この学習データを使用して学習を行った省略検出器 166、照応詞検出器 168、及び先行詞特定器 170 を用いてテキストの照応・省略解析を効率的に行うことが可能になる。また、学習データが大量になるため、これら検出器の精度が向上することが期待でき、高い精度でほぼ自動的にテキストの照応・省略解析を行うことができる。

[0110] なお、上記実施の形態では、質問応答システム 206 は単数であるものとして説明した。しかし本発明はそのような実施の形態には限定されない。質問応答システムが複数個利用できるのであれば、複数個の質問応答システムを利用してもよい。この場合、各質問応答システムがいずれも自然言語のテキストを入力とするのであれば、同一の質問文をそれらに送信するだけで回答候補を得ることができる。

[0111] [コンピュータによる実現]

上記実施の形態に係る学習データ生成補助装置 204 は、コンピュータハードウェアと、そのコンピュータハードウェア上で実行されるコンピュータプログラムとにより実現できる。図 19 はこのコンピュータシステム 930 の外観を示し、図 20 はコンピュータシステム 930 の内部構成を示す。

[0112] 図 19 を参照して、このコンピュータシステム 930 は、メモリポート 952 及び DVD (Digital Versatile Disc) ドライブ 950 を有するコンピュータ 940 と、キーボード 946 と、マウス 948 と、モニタ 942 とを含む。

[0113] 図 20 を参照して、コンピュータ 940 は、メモリポート 952 及び DVD ドライブ 950 に加えて、CPU (中央処理装置) 956 と、CPU 956、メモリポート 952 及び DVD ドライブ 950 に接続されたバス 966

と、ブートプログラム等を記憶する読出専用メモリ（ROM）958と、バス966に接続され、プログラム命令、システムプログラム及び作業データ等を記憶するランダムアクセスメモリ（RAM）960と、ハードディスク954を含む。コンピュータシステム930はさらに、他端末との通信を可能とするネットワーク968への接続を提供するネットワークインターフェイス（I/F）944を含む。

[0114] コンピュータシステム930を上記した実施の形態に係る学習データ生成補助装置204の各機能部として機能させるためのコンピュータプログラムは、DVDドライブ950又はメモリポート952に装着されるDVD962又はリムーバブルメモリ964に記憶され、さらにハードディスク954に転送される。又は、プログラムはネットワーク968を通じてコンピュータ940に送信されハードディスク954に記憶されてもよい。プログラムは実行の際にRAM960にロードされる。DVD962から、リムーバブルメモリ964から又はネットワーク968を介して、直接にRAM960にプログラムをロードしてもよい。

[0115] このプログラムは、コンピュータ940を、上記実施の形態に係る学習データ生成補助装置204の各機能部として機能させるための複数の命令からなる命令列を含む。コンピュータ940にこの動作を行わせるのに必要な基本的機能のいくつかはコンピュータ940上で動作するオペレーティングシステム若しくはサードパーティのプログラム又はコンピュータ940にインストールされる、ダイナミックリンク可能な各種プログラミングツールキット又はプログラムライブラリにより提供される。したがって、このプログラム自体はこの実施の形態のシステム及び方法を実現するのに必要な機能全てを必ずしも含まなくてよい。このプログラムは、命令の内、所望の結果が得られるように制御されたやり方で適切な機能又はプログラミングツールキット又はプログラムライブラリ内の適切なプログラムを実行時に動的に呼出すことにより、上記したシステムとしての機能を実現する命令のみを含んでいればよい。もちろん、プログラムのみで必要な機能を全て提供するようにし

てもよい。

[0116] 今回開示された実施の形態は単に例示であって、本発明が上記した実施の形態のみに制限されるわけではない。本発明の範囲は、発明の詳細な説明の記載を参酌した上で、請求の範囲の各請求項によって示され、そこに記載された文言と均等の意味及び範囲内での全ての変更を含む。

産業上の利用可能性

[0117] この発明は、自然言語処理に利用される学習データ等の整備を行う産業及びそのための装置、並びにそうして整備された学習データを用いて自然言語処理に関する多彩なサービスを提供する産業において利用可能である。

符号の説明

[0118] 150 照応・省略解析システム
160 アノテーションデータ生成補助システム
162 学習データDB
164 学習システム
166 省略検出器
168 照応詞検出器
170 先行詞特定器
172 照応・省略解析装置
190 既存小規模学習データDB
192 テキストアーカイブ
194, 228, 256 言語知識DB
196 言換え・カテゴリ辞書
198 パターン辞書
200, 220, 258 質問タイプDB
202 入出力装置
204 学習データ生成補助装置
206, 226, 260 質問応答システム
222 検出器学習装置

2 2 4, 2 6 2 選択制限 D B

2 5 2 照応・省略解析器

請求の範囲

- [請求項1] 自然言語文の照応・省略解析のためのアノテーションデータの作成を補助するアノテーション補助装置であって、
- 表示装置と入力装置とからなり、ユーザとの対話型処理によりユーザ入力を受け付ける入出力手段と、
- アノテーションの対象となるテキストアーカイブからテキストデータを読み出す読出手段と、
- 前記読出手段が読み出したテキストデータの形態素解析及び係り受け解析を行って、係り受け構造を示す情報が付された形態素列を出力する解析手段と、
- 前記解析手段の出力する形態素列中の述語を検索する述語検索手段と、
- 前記述語検索手段が検索した述語の各々が関与する係り受け関係において、当該述語に対して所定の関係にあるべき単語が省略されていること、又は指示語となっていることを検出し、アノテーション処理の対象として当該単語の位置を特定するための対象特定手段と、
- 前記対象特定手段が特定した前記単語の位置の各々について、当該位置に挿入されるべき表現の候補を、当該位置と周囲の形態素列との関係、及び言語知識を用いて推定する候補推定手段と、
- 前記候補推定手段により推定された候補を前記単語の位置と関連付けて記憶する候補記憶手段と、
- 前記アノテーション処理の対象の各々について、前記候補推定手段が推定した候補を前記候補記憶手段から読出し、ユーザがいずれかを選択可能なように前記表示装置に表示する候補表示手段と、
- 前記候補表示手段により表示された候補のいずれかを選択するユーザの指示に応答して、選択された候補を前記位置にアノテーションとして付加する対話的選択手段とを含む、アノテーション補助装置。
- [請求項2] 前記候補推定手段は、

前記対象特定手段が特定した前記単語の位置の各々について、当該単語の位置に入るべき単語を問い合わせる文を、当該位置の周囲の単語列を用いて生成する質問文生成手段と、

前記質問文生成手段が生成した質問文を、予め準備された質問応答システムに入力として与える質問文入力手段と、

前記質問文入力手段から与えられた質問文に対して前記質問応答システムから得られる回答から、前記単語の位置に挿入されるべき単語の候補を当該単語の位置と関連付けて前記候補記憶手段に記憶させる手段とを含む、請求項1に記載のアノテーション補助装置。

[請求項3]

前記候補推定手段は、さらに、

複数の言語知識規則を記憶する言語知識記憶手段と、

前記対象特定手段が特定した前記単語の位置の各々について、当該単語の位置と、当該単語の位置を特定する際に用いられた前記述語とを含む表現であって、前記言語知識記憶手段に記憶された言語知識のいずれかに適合する表現を特定する適合表現抽出手段と、

前記適合表現抽出手段により抽出された表現と、当該表現が適合した言語知識規則とを照合することにより、前記表現中の、前記対象特定手段が特定した前記単語の位置に挿入されるべき表現を、当該単語の位置に挿入されるべき文字列の候補として前記候補記憶手段に記憶させる手段とを含む、請求項2に記載のアノテーション補助装置。

[請求項4]

前記候補推定手段はさらに、

予め準備された既存のアノテーション済テキストデータベースに含まれるテキスト中で、省略又は照応関係に関するアノテーションが付されている箇所を検索するアノテーション検索手段と、

前記アノテーション検索手段により検索された箇所の各々について、各アノテーションごとに予め定められた方式にしたがって文を変形し、当該変形箇所を前記検索された箇所に対するアノテーションの候補として前記候補記憶手段に記憶させる手段とを含む、請求項3に記載

載のアノテーション補助装置。

[請求項5]

前記候補推定手段はさらに、

前記解析手段の出力する解析後の形態素列を、他の既存の照応・省略解析器への入力として与える手段と、

前記既存の照応・省略解析器による照応・省略解析結果を受け、当該照応・省略解析結果を、前記候補記憶手段に記憶させる手段とを含む、請求項2～請求項4のいずれかに記載のアノテーション補助装置。

[請求項6]

自然言語文の照応・省略解析のためのアノテーションデータの作成を補助するアノテーション補助装置としてコンピュータを動作させるコンピュータプログラムであって、コンピュータを、

当該コンピュータの表示装置と入力装置とを用いたユーザとの対話型処理によりユーザ入力を受け付ける入出力手段と、

アノテーションの対象となるテキストアーカイブからテキストデータを読み出す読出手段と、

前記読出手段が読み出したテキストデータの形態素解析及び係り受け解析を行って、係り受け構造を示す情報が付された形態素列を出力する解析手段と、

前記解析手段の出力する形態素列中の述語を検索する述語検索手段と、

前記述語検索手段が検索した述語の各々が関与する係り受け関係において、当該述語に対して所定の関係にあるべき単語が省略されていること、又は指示語となっていることを検出し、アノテーション処理の対象として当該単語の位置を特定するための対象特定手段と、

前記対象特定手段が特定した前記単語の位置の各々について、当該位置に挿入されるべき表現の候補を、当該位置と周囲の形態素列との関係、及び言語知識を用いて推定する候補推定手段と、

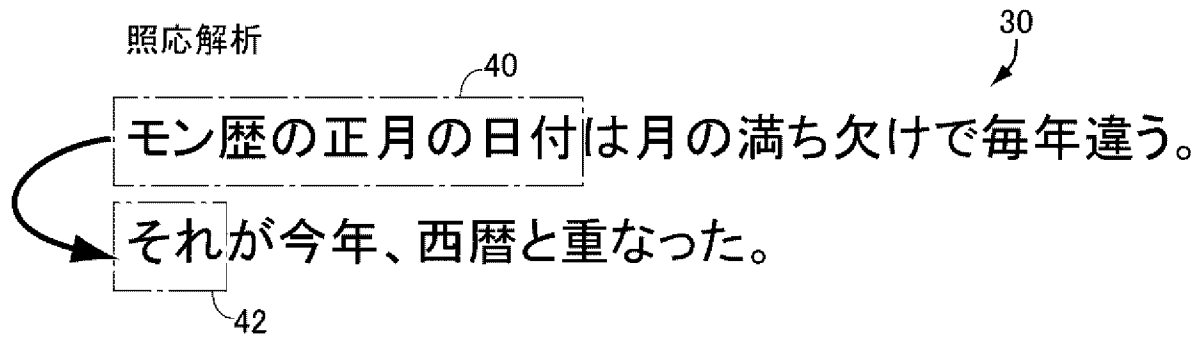
前記候補推定手段により推定された候補を前記単語の位置と関連付

けて記憶する候補記憶手段と、

前記アノテーション処理の対象の各々について、前記候補推定手段が推定した候補を前記候補記憶手段から読出し、ユーザがいずれかを選択可能なように前記表示装置に表示する候補表示手段と、

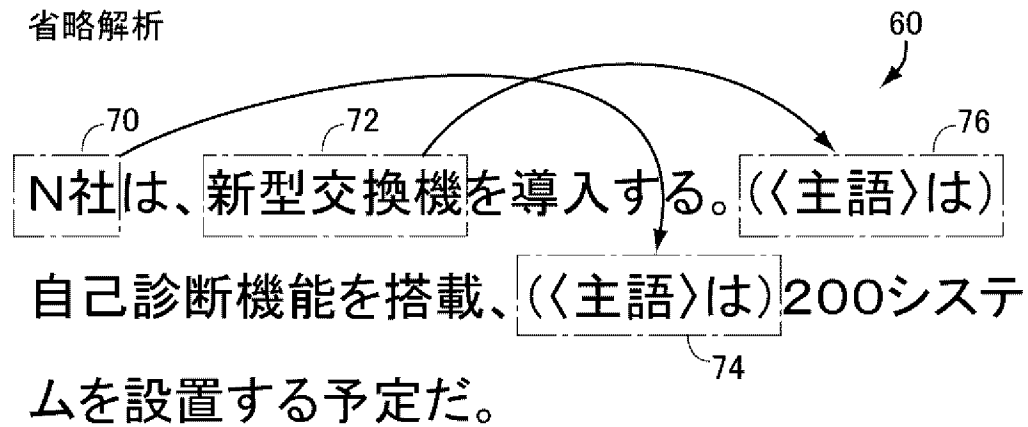
前記候補表示手段により表示された候補のいずれかを選択するユーザの指示に応答して、選択された候補を前記位置にアノテーションとして付加する対話的選択手段として機能させる、コンピュータプログラム。

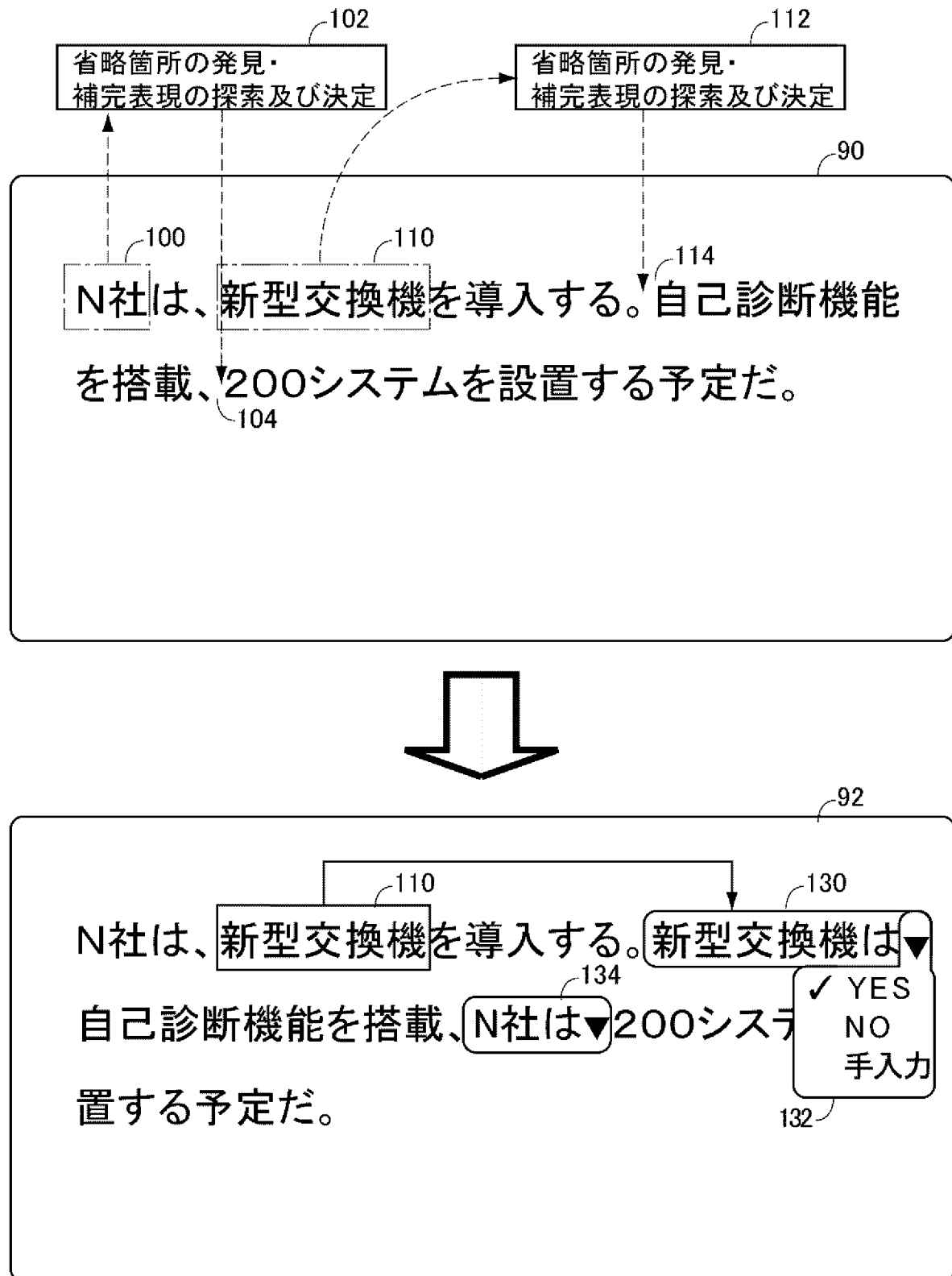
[図1]



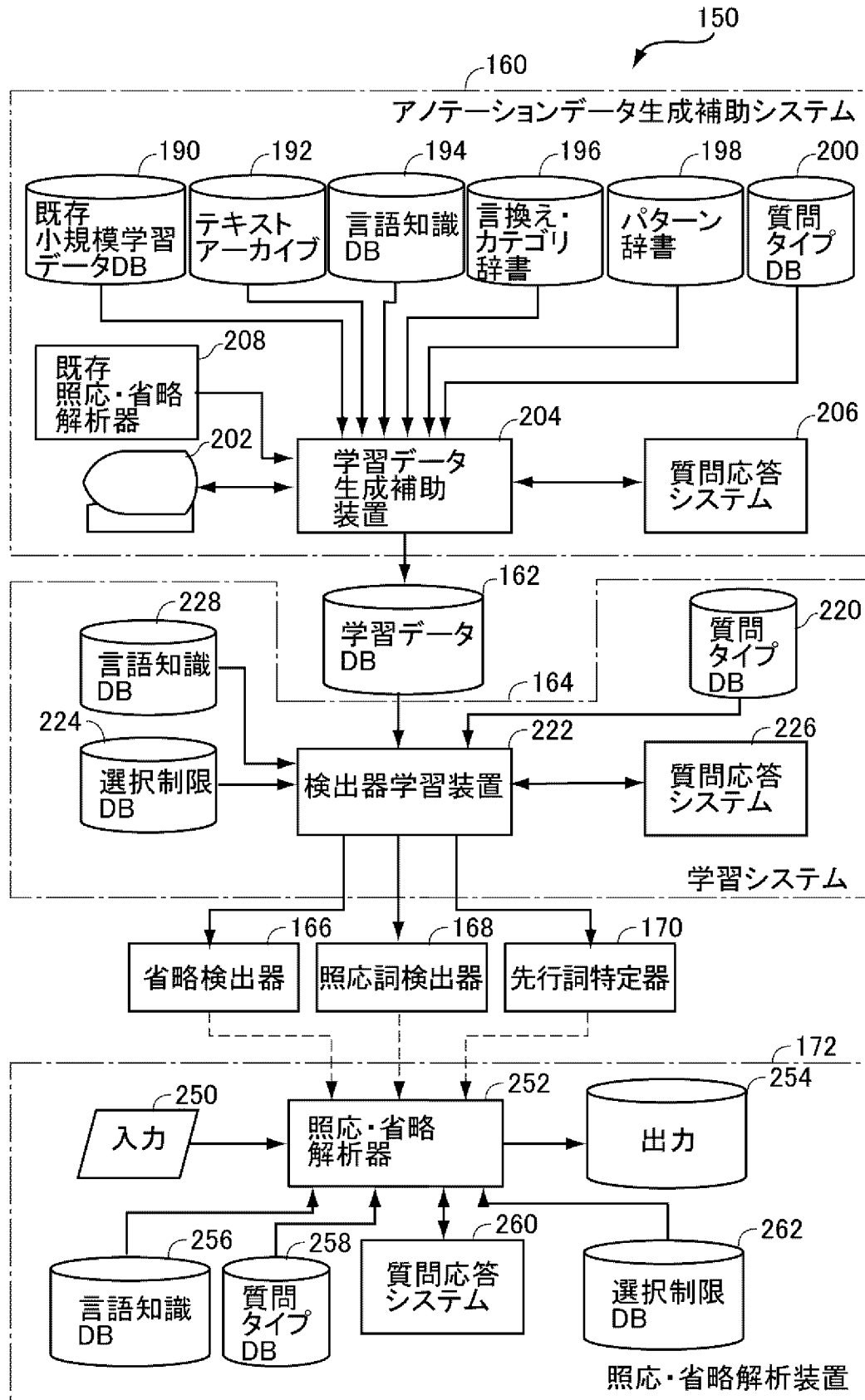
[図2]

省略解析





[図4]



[図5]

194

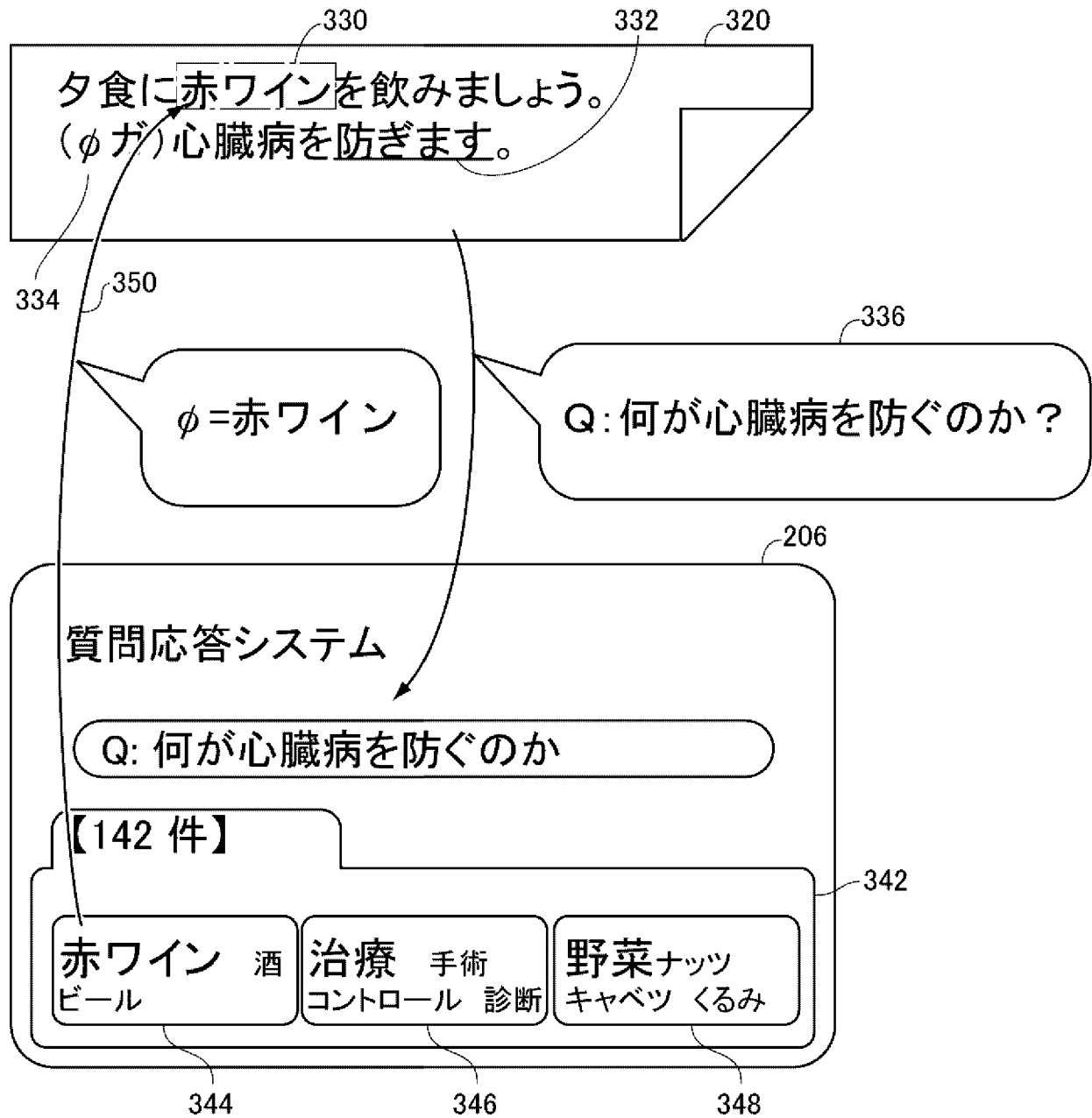
<u>因果関係知識</u>	Xがタバコを吸う ...	→	Xが肺がんの可能性を高める ...	290
<u>含意関係知識</u>	XがYの著者だ ...	→	XがYを執筆した ...	292
<u>事態生起順序関係知識</u>	XがYを発見する XがYを煮る ...	→ →	XがYを発表する ZがYを食べる(X≡Z) ...	294
<u>事態生起無効化関係知識</u>	Xを禁止する ...	→	Xを公開できない	296

[図6]

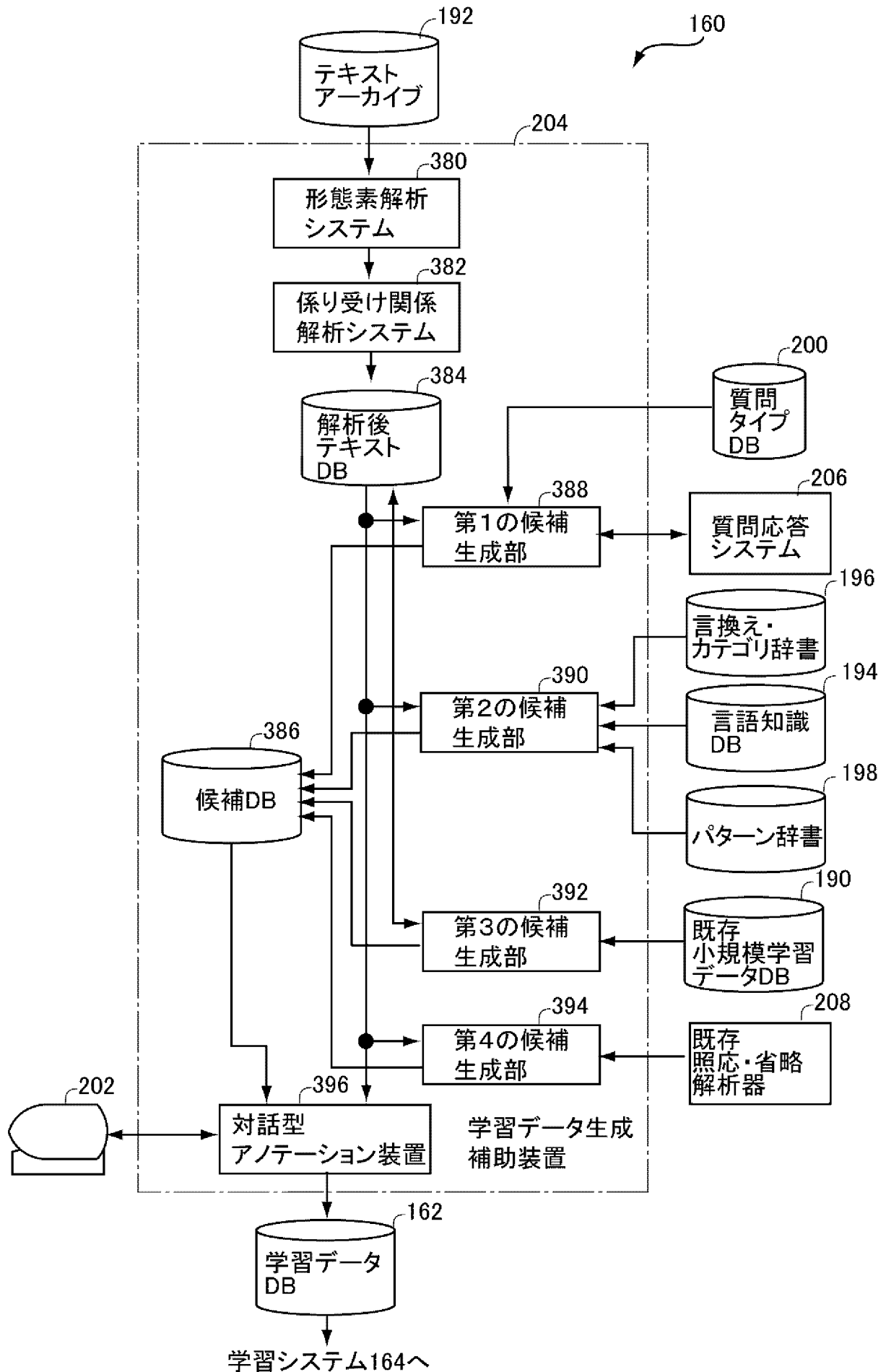
200, 220, 258

動詞	文法役割	質問タイプ
...	...	...
食べる	主語	誰が
食べる	目的語	何を
...	...	...
予防する	主語	何が
予防する	目的語	何を
...	...	...

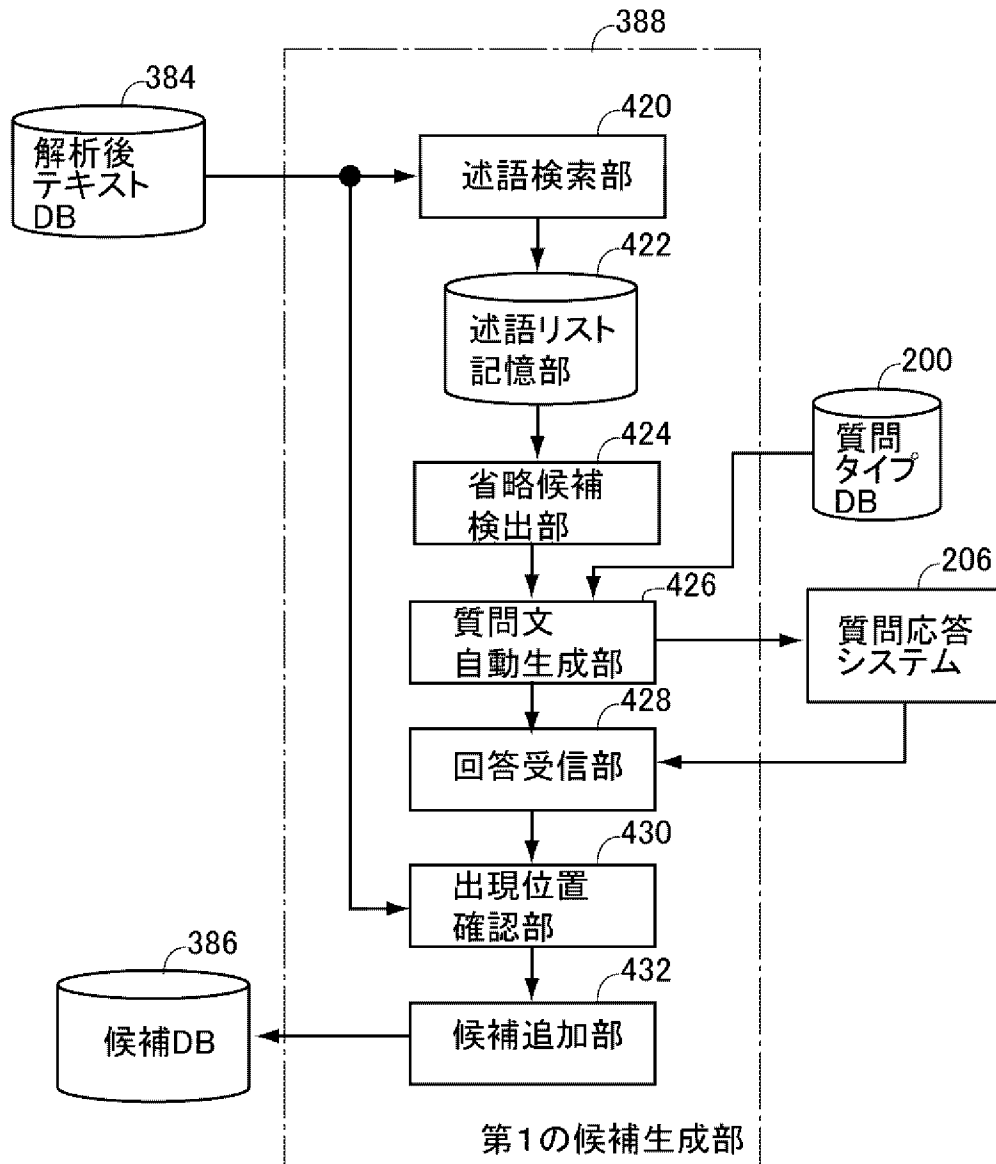
[図7]



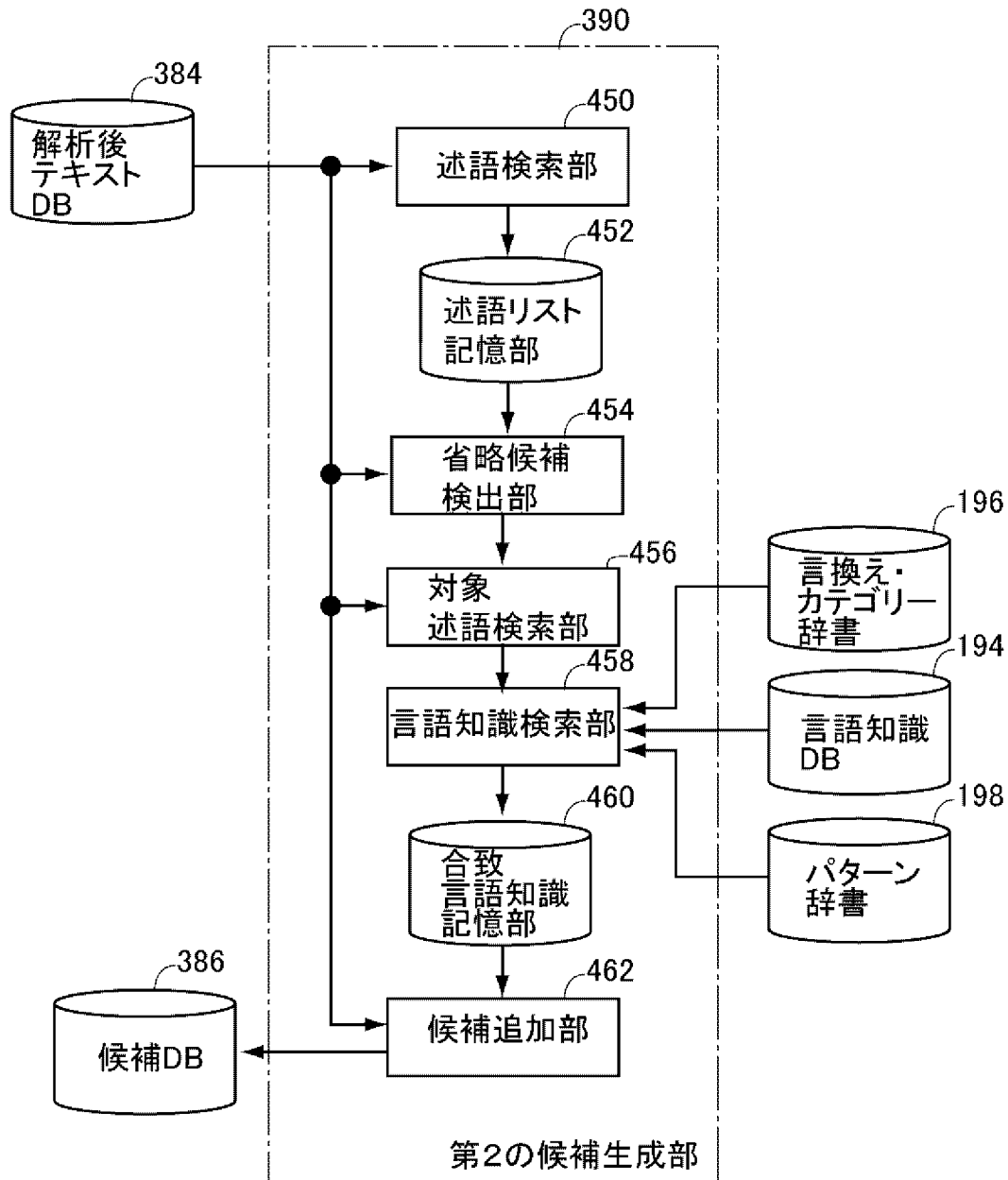
[図8]



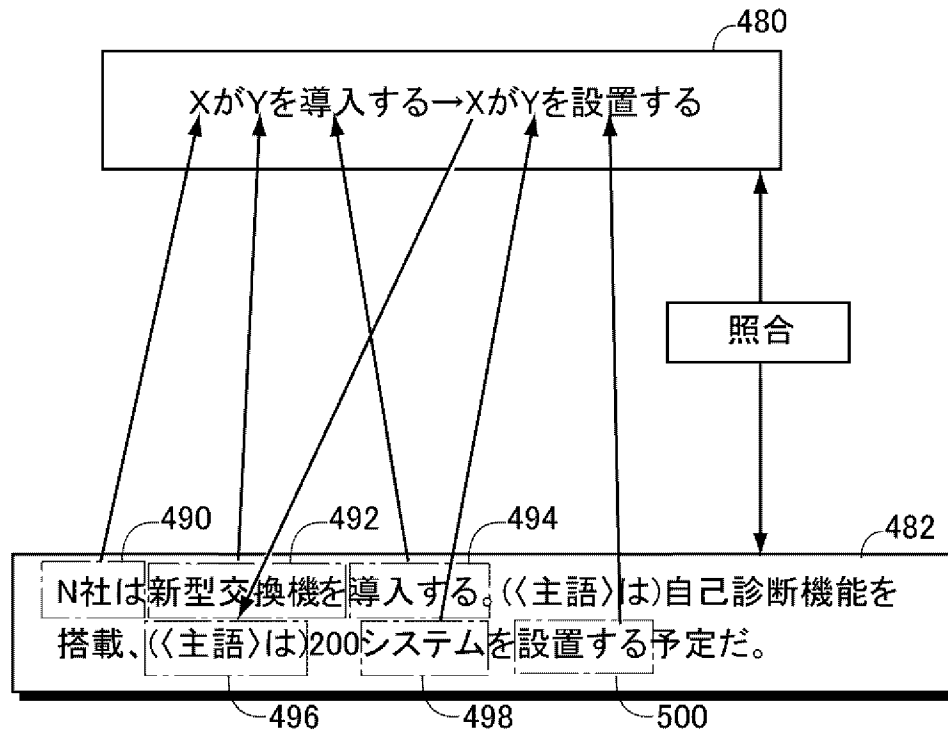
[図9]



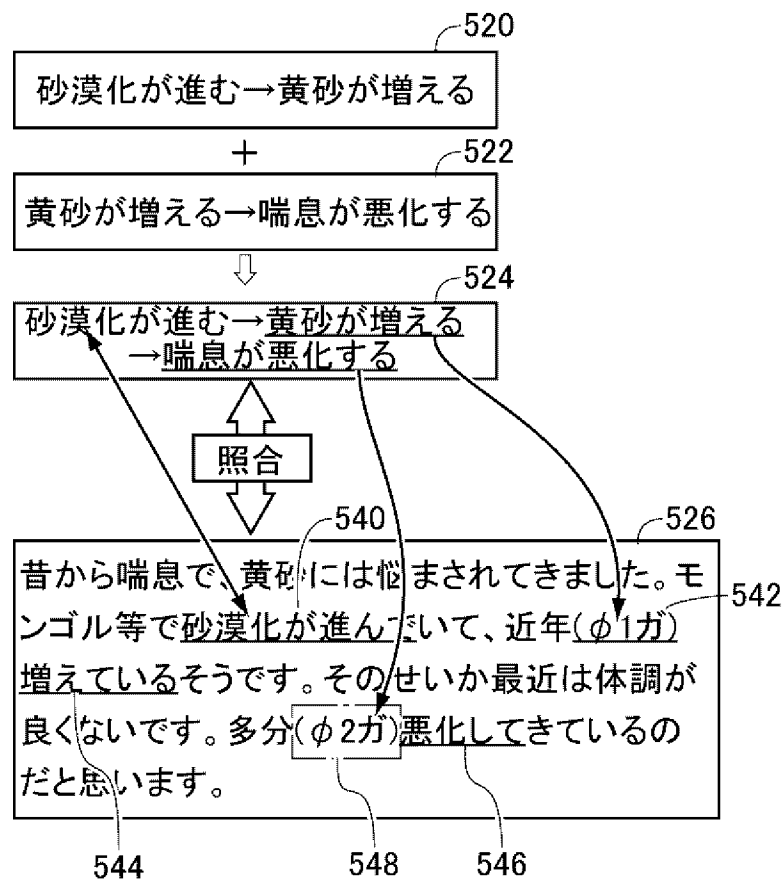
[図10]



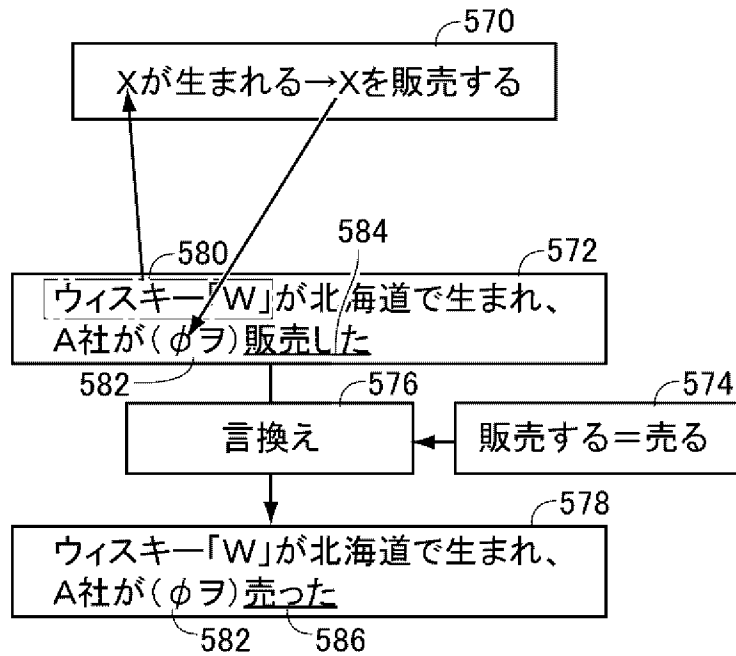
[図11]



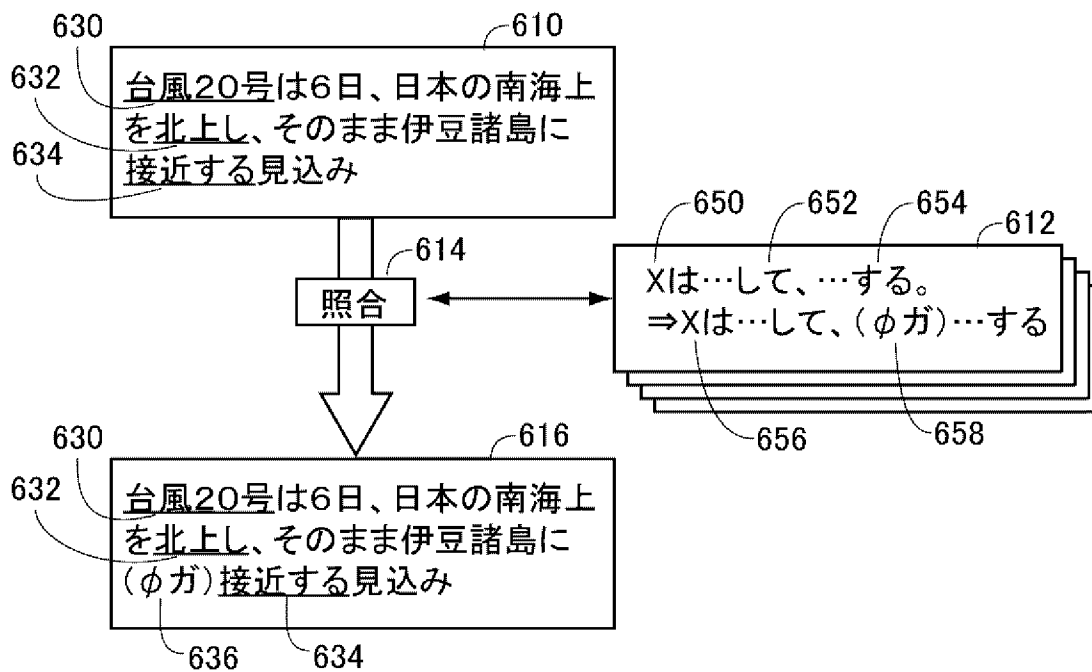
[図12]



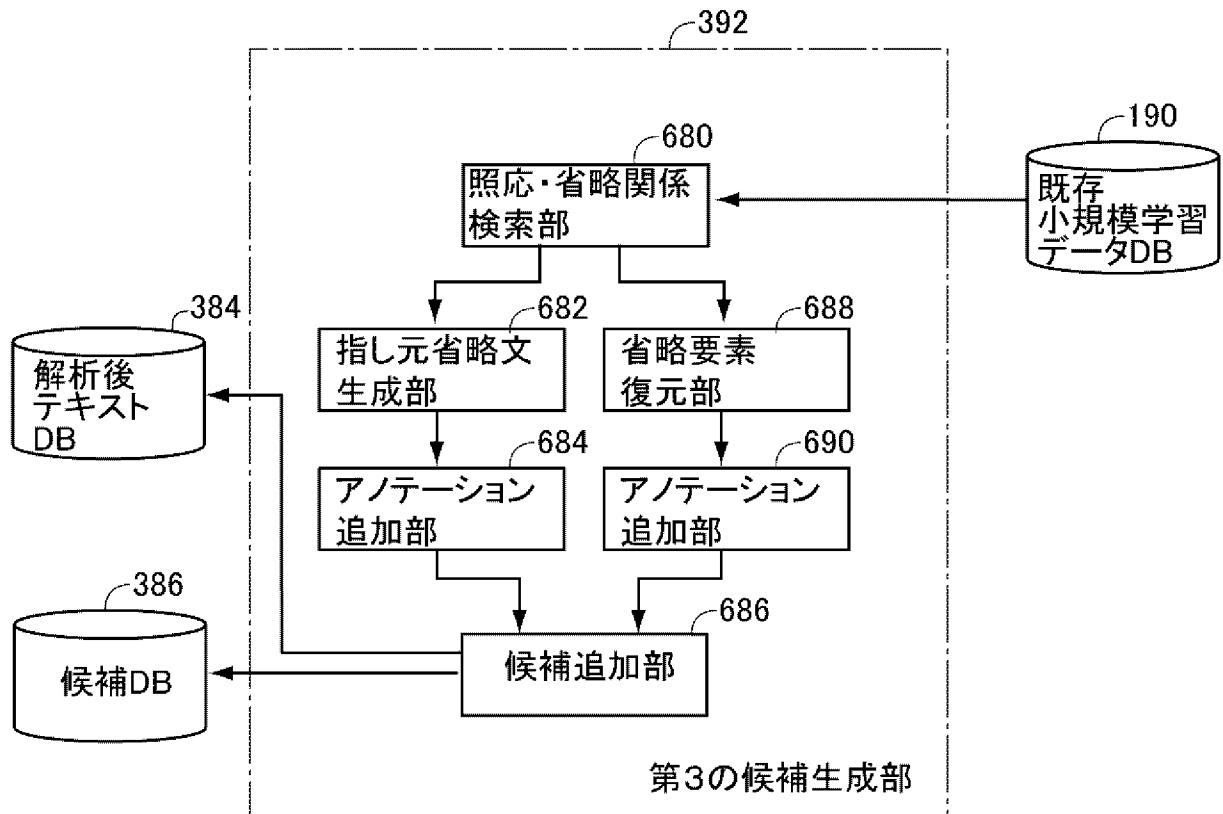
[図13]



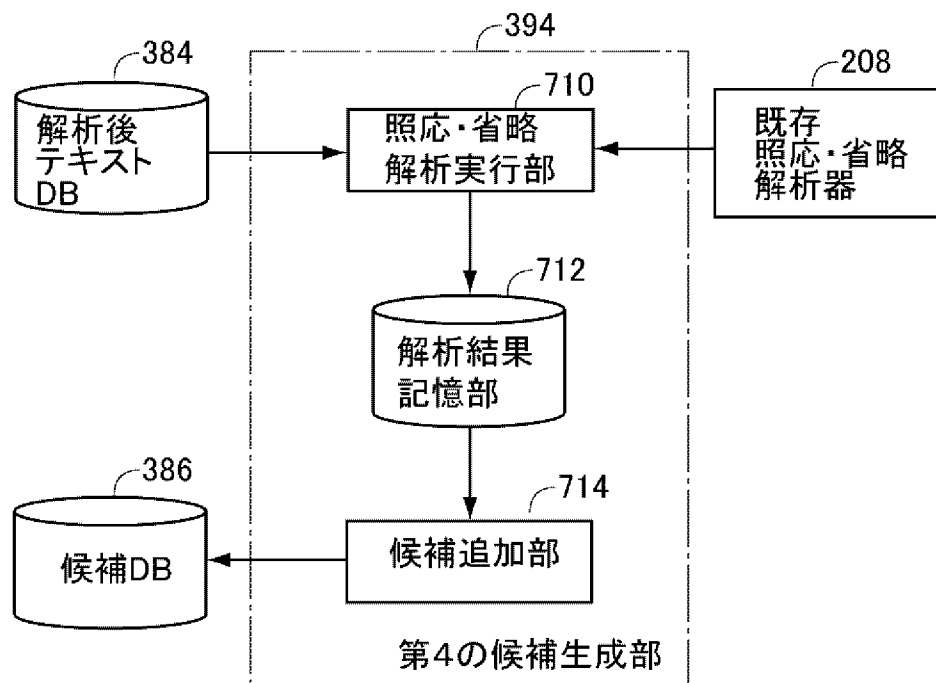
[図14]



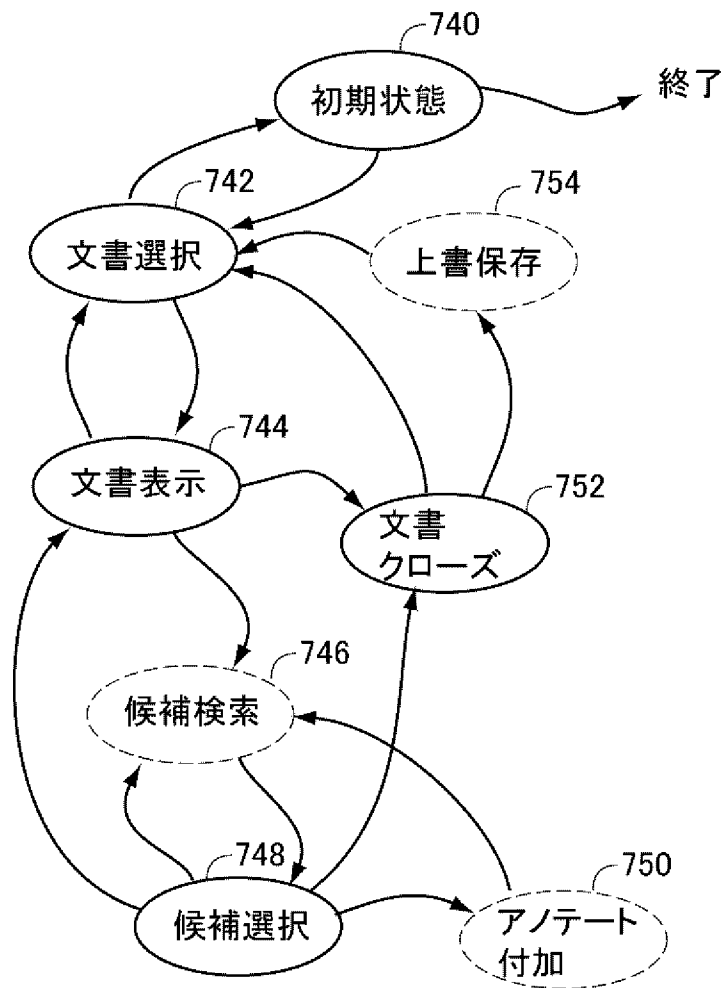
[図15]



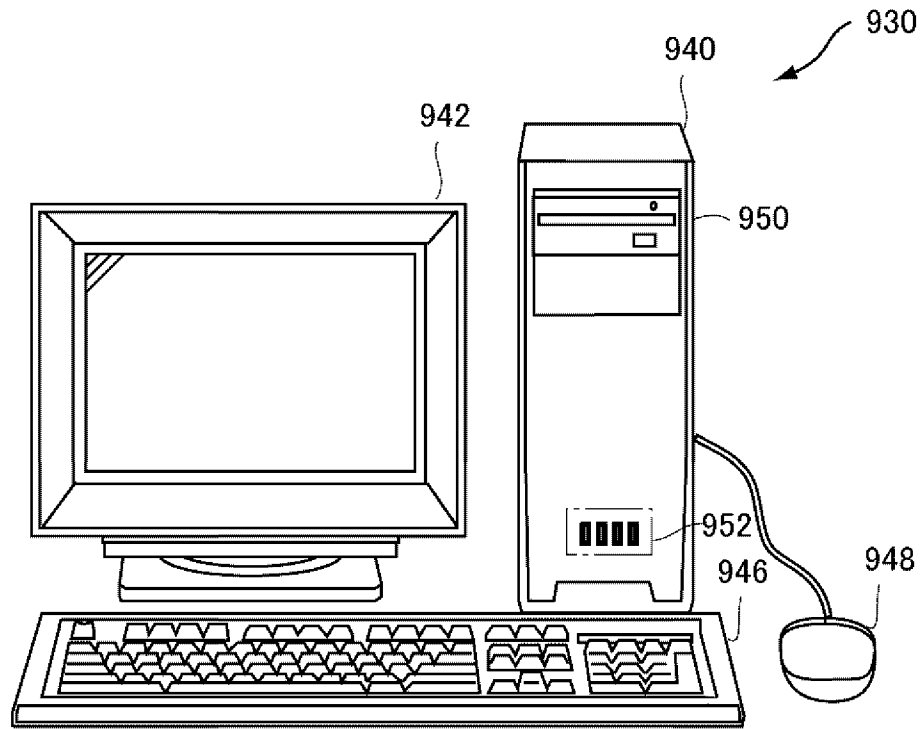
[図16]



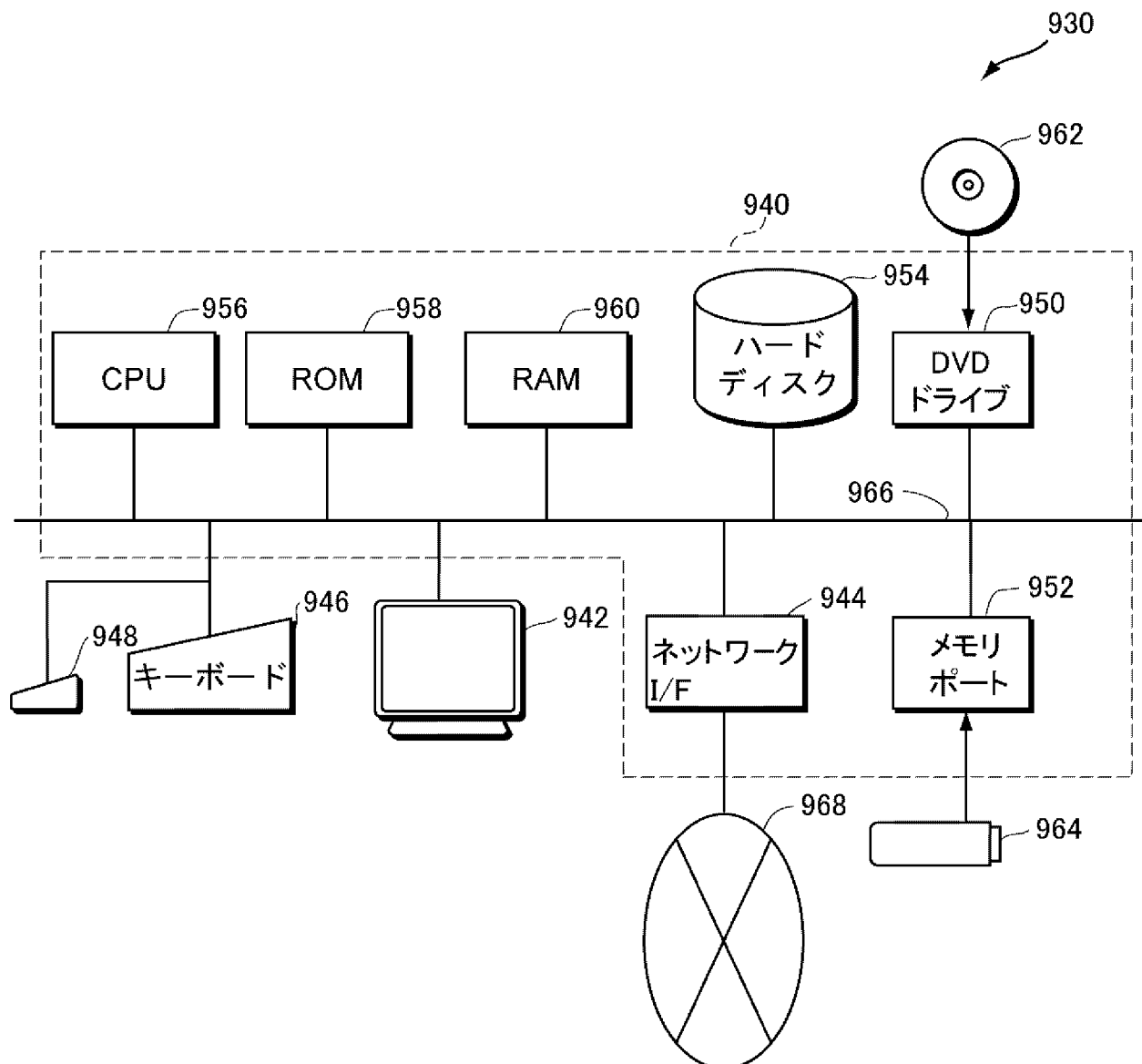
[図17]



[図19]



[図20]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2016/051577

A. CLASSIFICATION OF SUBJECT MATTER

G06F17/27(2006.01) i, G06F17/24(2006.01) i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F17/27, G06F17/24

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2016
Kokai Jitsuyo Shinan Koho	1971-2016	Toroku Jitsuyo Shinan Koho	1994-2016

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	JP 2004-5648 A (Microsoft Corp.), 08 January 2004 (08.01.2004), paragraphs [0001] to [0014], [0028], [0030], [0036], [0039] to [0040], [0047] to [0048], [0053], [0056] to [0057]; fig. 2, 3 & US 2003/0212544 A1 & US 2009/0276380 A1 & EP 1361522 A2 & CN 1457041 A & AT 519164 T & ES 2368213 T3	1-6
Y	Shinpei KAWAGUCHI, Tatsunori MORI, "Shoo Kaiseki ni Chishikigen o Riyo shita Kanren Shitsumongun ni Taisuru Shitsumon Oto", Proceedings of the 12th Annual Meeting of the Association for Natural Language Processing, 13 March 2006 (13.03.2006), pages 927 to 930	1-6

☒ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
24 February 2016 (24.02.16)

Date of mailing of the international search report
08 March 2016 (08.03.16)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2016/051577

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Kazushi IKEDA et al., "Estimation and Complementation Approach for the Analysis of the Omission of Postposition on Colloquial Style Sentences", IPSJ SIG Notes Heisei 22 Nendo [CD-ROM], 15 December 2010 (15.12.2010), pages 1 to 8	2-5
Y	Takafumi SUZUKI et al., "Daihyo·Hasei Kankei o Riyo shita Nihongo Kino Hyogen no Kaiseki Hoshiki no Hyoka", The Association for Natural Language Processing Dai 18 Kai Nenji Taikai Happyo Ronbunshu Tutorial Honkaigi [CD-ROM], 13 March 2012 (13.03.2012), pages 598 to 601	4-5
A	Yuji MATSUMOTO, "Corpus Annotation and Annotation-Assistance Tools", Journal of Japanese Society for Artificial Intelligence, 01 September 2009 (01.09.2009), vol.24, no.5, pages 632 to 639	1-6

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G06F17/27(2006.01)i, G06F17/24(2006.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int.Cl. G06F17/27, G06F17/24

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2016年
日本国実用新案登録公報	1996-2016年
日本国登録実用新案公報	1994-2016年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
Y	JP 2004-5648 A（マイクロソフト コーポレーション）2004.01.08, 段落 0001-0014, 0028, 0030, 0036, 0039-0040, 0047-0048, 0053, 0056-0057、 図 2, 3 & US 2003/0212544 A1 & US 2009/0276380 A1 & EP 1361522 A2 & CN 1457041 A & AT 519164 T & ES 2368213 T3	1-6
Y	川口 晋平、森 辰則、照応解析に知識源を利用した関連質問群に 対する質問応答、言語処理学会第12回年次大会発表論文集、 2006.03.13, p.927-930	1-6

C欄の続きにも文献が列挙されている。

パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの
「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの
「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）
「O」口頭による開示、使用、展示等に言及する文献
「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの
「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの
「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの
「&」同一パテントファミリー文献

国際調査を完了した日

24.02.2016

国際調査報告の発送日

08.03.2016

国際調査機関の名称及びあて先

日本国特許庁（ISA/J P）

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

長 由紀子

5M

4233

電話番号 03-3581-1101 内線 3599

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
Y	池田 和史 外4名, 口語文書の解析精度向上のための助詞落ち推定および補完手法の提案, 情報処理学会研究報告 平成22年度 [CD-ROM], 2010.12.15, p.1-8	2-5
Y	鈴木 敬文 外4名, 代表・派生関係を利用した日本語機能表現の解析方式の評価, 言語処理学会第18回年次大会発表論文集 チュートリアル 本会議 [CD-ROM], 2012.03.13, p.598-601	4-5
A	松本 裕治, コーパスへの自動アノテーションツールとアノテーション支援環境の構築, 人工知能学会誌, 2009.09.01, 第24巻第5号, p.632-639	1-6