(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2020-52902
(P2020-52902A)
(43)公開日　令和2年4月2日(2020.4.2)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 40/289	(2020.01)	G06F	17/27	680	5B091
G06F 16/30	(2019.01)	G06F	17/30	170A	
G06F 16/00	(2019.01)	G06F	17/30	220Z	
G06F 40/242	(2020.01)	G06F	17/27	635	

審査請求 未請求　請求項の数 7　ＯＬ　（全 17 頁）

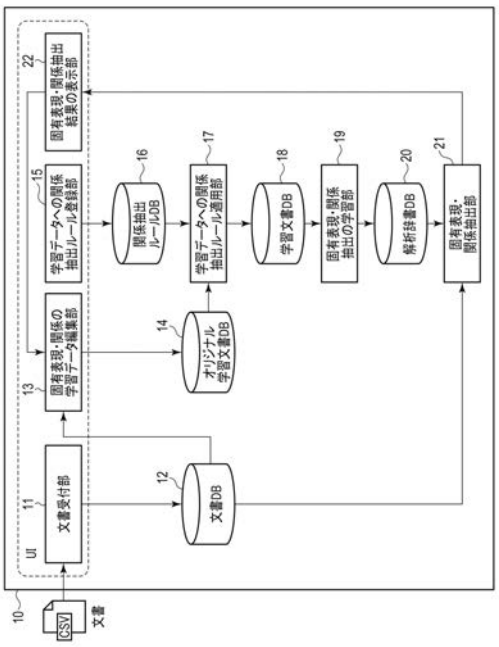
(21)出願番号　特願2018-183861 (P2018-183861)
(22)出願日　平成30年9月28日 (2018.9.28)

(71)出願人　000003078
株式会社東芝
東京都港区芝浦一丁目１番１号
(71)出願人　301063496
東芝デジタルソリューションズ株式会社
神奈川県川崎市幸区堀川町７２番地３４
(74)代理人　100108855
弁理士　蔵田　昌俊
(74)代理人　100103034
弁理士　野河　信久
(74)代理人　100075672
弁理士　峰　隆司
(74)代理人　100153051
弁理士　河野　直樹

最終頁に続く

(54)【発明の名称】固有表現抽出装置、方法およびプログラム

(57)【要約】
【課題】文書からの固有表現抽出の精度を向上させる。
【解決手段】実施形態における固有表現抽出装置は、文書データから固有表現および固有表現同士の関係を抽出する抽出ルールを定めた抽出用辞書を格納する手段と、抽出対象である抽出用文書データおよび学習用文書データの入力を受け付ける文書受付手段と、抽出用辞書を用いて抽出用文書データから固有表現および関係を抽出する抽出手段と、入力操作に従い、学習用文書データにおける文字列のうち抽出させる固有表現に対応する文字列を指定する指定手段と、抽出用文書データから抽出させる、固有表現の分類同士の関係を定めた関係抽出ルールを格納する手段と、関係抽出ルールを適用することで、指定された固有表現のうち関係抽出ルールの分類に属する固有表現同士の関係を設定した学習文書を生成する生成手段と、学習文書に基づいて抽出用辞書を学習する学習手段とを有する。
【選択図】　図１



【特許請求の範囲】
【請求項１】
文書データから当該文書データの固有表現および固有表現同士の関係を抽出する抽出ルールを定めた抽出用辞書を格納する第１の格納手段と、
前記固有表現および前記関係の抽出対象である抽出用文書データおよび前記抽出用辞書の学習に用いられる学習用文書データの入力を受け付ける文書受付手段と、
前記抽出用辞書を用いて、前記文書受付手段により受け付けた抽出用文書データから固有表現および固有表現同士の関係をそれぞれ抽出する抽出手段と、
入力操作に従い、前記文書受付手段により受け付けた学習用文書データにおける文字列のうち前記抽出手段により抽出させる固有表現に対応する文字列を指定する指定手段と、
前記抽出用文書データから抽出させる、固有表現の分類同士の関係を定めた関係抽出ルールを格納する第２の格納手段と、
前記第２の格納手段に格納された関係抽出ルールを適用することで、前記指定手段により指定された固有表現のうち前記関係抽出ルールで定められた分類に属する固有表現同士の関係を設定した学習文書を生成する生成手段と、
前記生成手段により生成された学習文書に基づいて、前記抽出用辞書を学習する学習手段と、
を備えた固有表現抽出装置。
【請求項２】
前記抽出手段により抽出された固有表現を出力する第１の出力手段と、
前記第１の出力手段により出力された固有表現のうち入力操作で指定された固有表現の抽出元の文書データを出力する第２の出力手段と、
入力操作に従い、前記第２の出力手段により出力された抽出元の文書データの固有表現を編集する編集手段と、
をさらに備えた請求項１に記載の固有表現抽出装置。
【請求項３】
前記抽出手段により抽出された固有表現同士の関係を出力する第１の出力手段と、
前記第１の出力手段により表示された固有表現同士の関係のうち入力操作で指定された固有表現同士の関係の抽出元の文書データを出力する第２の出力手段と、
入力操作に従い、前記第２の出力手段により出力された抽出元の文書データの固有表現同士の関係を編集する編集手段と、
をさらに備えた請求項１に記載の固有表現抽出装置。
【請求項４】
文書データから当該文書データの固有表現および固有表現同士の関係を抽出する抽出ルールを定めた抽出用辞書を格納する格納手段と、
前記固有表現および前記関係の抽出対象である抽出用文書データおよび前記抽出用辞書の学習に用いられる学習用文書データの入力を受け付ける文書受付手段と、
前記抽出用辞書を用いて、前記文書受付手段により受け付けた抽出用文書データから固有表現および固有表現同士の関係をそれぞれ抽出する抽出手段と、
入力操作に従い、前記文書受付手段により受け付けた学習用文書データにおける文字列のうち前記抽出手段により抽出させる固有表現に対応する文字列および固有表現同士の関係をそれぞれ指定する指定手段と、
前記指定手段により指定された固有表現に対応する文字列および固有表現同士の関係に基づいて、前記抽出用辞書を学習する学習手段と、
前記指定手段により指定された固有表現に対応する文字列および固有表現同士の関係のうち、前記抽出手段により抽出されなかった、固有表現に対応する文字列および固有表現同士の関係を出力する出力手段と、
を備えた固有表現抽出装置。
【請求項５】
文書データから当該文書データの固有表現および固有表現同士の関係を抽出する抽出ル

ールを定めた抽出用辞書を格納する格納手段と、
前記固有表現および前記関係の抽出対象である抽出用文書データおよび前記抽出用辞書の学習に用いられる学習用文書データの入力を受け付ける文書受付手段と、
前記抽出用辞書を用いて、前記文書受付手段により受け付けた抽出用文書データから固有表現および固有表現同士の関係をそれぞれ抽出する抽出手段と、
入力操作に従い、前記文書受付手段により受け付けた学習用文書データにおける文字列のうち前記抽出手段により抽出させる固有表現に対応する文字列および固有表現同士の関係をそれぞれ指定する指定手段と、
前記指定手段により指定された固有表現に対応する文字列および固有表現同士の関係に基づいて、前記抽出用辞書を学習する学習手段と、
前記指定手段により指定されない固有表現に対応する文字列および固有表現同士の関係のうち、前記抽出手段により抽出された、固有表現に対応する文字列および固有表現同士の関係を出力する出力手段と、
を備えた固有表現抽出装置。
【請求項６】
固有表現抽出装置に適用する方法であって、
固有表現および固有表現同士の関係の抽出対象である抽出用文書データおよび、前記抽出用文書データから当該抽出用文書データの固有表現および固有表現同士の関係を抽出する抽出ルールを定めた抽出用辞書の学習に用いられる学習用文書データの入力を受け付ける処理と、
文書データから当該文書データの固有表現および固有表現同士の関係を抽出する抽出ルールを定めた抽出用辞書を用いて、前記受け付けた抽出用文書データから固有表現および固有表現同士の関係をそれぞれ抽出する処理と、
入力操作に従い、前記受け付けた学習用文書データにおける文字列のうち前記抽出させる固有表現に対応する文字列を指定する処理と、
前記抽出用文書データから抽出させる、固有表現の分類同士の関係を定めた関係抽出ルールを適用することで、前記指定された固有表現のうち前記関係抽出ルールで定められた分類に属する固有表現同士の関係を設定した学習文書を生成する処理と、
前記生成された学習文書に基づいて、前記抽出用辞書を学習する処理と、
を実行する固有表現抽出方法。
【請求項７】
コンピュータを、
文書データから当該文書データの固有表現および固有表現同士の関係を抽出する抽出ルールを定めた抽出用辞書を格納する第１の格納手段、
前記固有表現および前記関係の抽出対象である抽出用文書データおよび前記抽出用辞書の学習に用いられる学習用文書データの入力を受け付ける文書受付手段、
前記抽出用辞書を用いて、前記文書受付手段により受け付けた抽出用文書データから固有表現および固有表現同士の関係をそれぞれ抽出する抽出手段、
入力操作に従い、前記文書受付手段により受け付けた学習用文書データにおける文字列のうち前記抽出手段により抽出させる固有表現に対応する文字列を指定する指定手段、
前記抽出用文書データから抽出させる、固有表現の分類同士の関係を定めた関係抽出ルールを格納する第２の格納手段、
前記第２の格納手段に格納された関係抽出ルールを適用することで、前記指定手段により指定された固有表現のうち前記関係抽出ルールで定められた分類に属する固有表現同士の関係を設定した学習文書を生成する生成手段、および
前記生成手段により生成された学習文書に基づいて、前記抽出用辞書を学習する学習手段、
として機能させる固有表現抽出処理プログラム。
【発明の詳細な説明】
【技術分野】

【０００１】
本発明の実施形態は、固有表現抽出装置、方法およびプログラムに関する。
【背景技術】
【０００２】
従来、人手によるルール又は機械学習などの様々な手法によって、文書データ中に出現する固有表現を抽出する仕組みが提案されてきた。
【０００３】
また、文書データから抽出した固有表現について、この固有表現の分類名が出現する度合いから、分類名の重みを算出することで、どの固有表現を出力するかを判定するなどの応用技術も存在する。
【先行技術文献】
【特許文献】
【０００４】
【特許文献１】特開２００７－１４８７８５号公報
【発明の概要】
【発明が解決しようとする課題】
【０００５】
しかしながら、他の認識技術と同様に、固有表現抽出にも、理想的には１００％の精度が期待されており、さらなる精度向上が求められている。
【０００６】
本発明が解決しようとする課題は、文書からの固有表現抽出の精度を向上させることが可能な固有表現抽出装置、方法および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】
【０００７】
実施形態に係る固有表現抽出装置は、文書データから当該文書データの固有表現および固有表現同士の関係を抽出する抽出ルールを定めた抽出用辞書を格納する第１の格納手段と、前記固有表現および前記関係の抽出対象である抽出用文書データおよび抽出用辞書の学習に用いられる学習用文書データの入力を受け付ける文書受付手段と、前記抽出用辞書を用いて、前記文書受付手段により受け付けた抽出用文書データから固有表現および固有表現同士の関係をそれぞれ抽出する抽出手段と、入力操作に従い、前記文書受付手段により受け付けた学習用文書データにおける文字列のうち前記抽出手段により抽出させる固有表現に対応する文字列を指定する指定手段と、抽出用文書データから抽出させる、固有表現の分類同士の関係を定めた関係抽出ルールを格納する第２の格納手段と、前記第２の格納手段に格納された関係抽出ルールを適用することで、前記指定手段により指定された固有表現のうち前記関係抽出ルールで定められた分類に属する固有表現同士の関係を設定した学習文書を生成する生成手段と、前記生成手段により生成された学習文書に基づいて、前記抽出用辞書を学習する学習手段とを有する。
【発明の効果】
【０００８】
本発明によれば、文書からの固有表現抽出の精度を向上させることができる。
【図面の簡単な説明】
【０００９】
【図１】実施形態に係る固有表現抽出装置の機能構成例を示すブロック図。
【図２】実施形態に係る固有表現抽出装置の文書ＤＢに格納される文書データの一例を表形式で示す図。
【図３】実施形態に係る固有表現抽出装置のオリジナル学習文書ＤＢに格納される、固有表現に係る学習データの一例を表形式で示す図。
【図４】実施形態に係る固有表現抽出装置のオリジナル学習文書ＤＢに格納される、固有表現同士の関係に係る学習データの一例を表形式で示す図。
【図５】実施形態に係る固有表現抽出装置の関係抽出ルールＤＢに格納される関係抽出ル

ールの一例を表形式で示す図。
【図６】実施形態に係る固有表現抽出装置の解析辞書ＤＢに格納される解析辞書の一例を表形式で示す図。
【図７】実施形態に係る固有表現抽出装置の第１の処理手順の一例を示すフローチャート。
【図８】実施形態に係る固有表現抽出装置による、文書データの記述へのタグの付与時の表示画面の一例を示す図。
【図９】実施形態に係る固有表現抽出装置の第２の処理手順の一例を示すフローチャート。
【図１０】実施形態に係る固有表現抽出装置による、抽出されたタグ、およびタグ同士の関係の表示画面の一例を示す図。
【図１１】実施形態に係る固有表現抽出装置による、抽出されたタグ、およびタグ同士の関係の編集画面の一例を示す図。
【図１２】実施形態に係る固有表現抽出装置の第３の処理手順の一例を示すフローチャート。
【図１３】実施形態に係る固有表現抽出装置による、文書データのタグ同士の関係の付与時の表示画面の一例を示す図。
【図１４】実施形態に係る固有表現抽出装置による、抽出されなかったタグ、およびタグ同士の関係の表示画面の一例を示す図。
【図１５】実施形態に係る固有表現抽出装置の第４の処理手順の一例を示すフローチャート。
【図１６】実施形態に係る固有表現抽出装置による、学習されなかったが抽出されたタグ、およびタグ同士の関係の表示画面の一例を示す図。
【発明を実施するための形態】
【００１０】
以下、実施形態について図面を用いて説明する。
図１は、本発明の実施形態に係る固有表現抽出装置の機能構成例を示すブロック図である。
図１に示すように、実施形態に係る固有表現抽出装置１０は、文書受付部１１、文書ＤＢ（データベース）１２、固有表現・関係の学習データ編集部１３、オリジナル学習文書ＤＢ１４、学習データへの関係抽出ルール登録部１５、関係抽出ルールＤＢ１６、学習データへの関係抽出ルール適用部１７、学習文書ＤＢ１８、固有表現・関係抽出の学習部１９、解析辞書ＤＢ２０、固有表現・関係抽出部２１、および固有表現・関係抽出結果の表示部２２を有する。
【００１１】
また、固有表現抽出装置１０は、パーソナルコンピュータ（ＰＣ）などのコンピュータデバイスを用いたシステムにより実現可能である。例えば、コンピュータデバイスは、ＣＰＵ（Central Processing Unit）などのプロセッサと、プロセッサに接続されるメモリと、入出力インタフェースとを備える。このうちメモリは、不揮発性メモリなどの記憶媒体を有する記憶装置により構成される。
【００１２】
文書受付部１１、固有表現・関係の学習データ編集部１３、学習データへの関係抽出ルール登録部１５、学習データへの関係抽出ルール適用部１７、固有表現・関係抽出の学習部１９、固有表現・関係抽出部２１、および固有表現・関係抽出結果の表示部２２の機能は、例えば、プロセッサがメモリに格納されているプログラムを読み出して実行することにより実現される。なお、これらの機能の一部または全部は、特定用途向け集積回路（ＡＳＩＣ）などの回路によって実現されてもよい。
【００１３】
上記の機能のうち、文書受付部１１、固有表現・関係の学習データ編集部１３、学習データへの関係抽出ルール登録部１５、および固有表現・関係抽出結果の表示部２２の機能

は、ユーザインタフェース（ＵＩ）における図示しない入力装置および表示装置と協働した機能として実現することができる。入力装置は、例えばキーボードおよびマウスである。表示装置は、例えば液晶ディスプレイである。

【００１４】

文書ＤＢ１２、オリジナル学習文書ＤＢ１４、関係抽出ルールＤＢ１６、学習文書ＤＢ１８、解析辞書ＤＢ２０は、上記メモリのうち随時書込および読み出しが可能な不揮発性メモリに設けられる。

【００１５】

固有表現抽出装置１０は、文書データにおける固有表現（以下、タグと称することがある）の抽出結果と、固有表現同士の関係（以下、リンクと称することがある）の抽出結果とをあわせて表示装置に表示することができる。また、ユーザに、表示を参照して固有表現および固有表現同士の関係の誤抽出および未抽出を発見させることで、文書データからの固有表現および固有表現同士の関係抽出ルールを定めた解析辞書（抽出用辞書と称することもある）の学習に用いられる学習データを修正することを支援することもできる。

【００１６】

文書受付部１１は、１つ以上の文書データの入力（登録）を受け付けて、この受け付けた文書データを文書ＤＢ１２に格納する。この格納される文書データは、（１）固有表現、および固有表現同士の関係抽出の対象となる抽出用文書データである場合と、（２）抽出用文書データからの固有表現、および固有表現同士の関係抽出ルールを定めた解析辞書の学習に用いる学習用文書データである場合とがある。

【００１７】

図２は、実施形態に係る固有表現抽出装置１０の文書ＤＢ１２に格納される文書データの一例を表形式で示す図である。

図２に示した例では、文書ＤＢ１２に格納される文書データは、（１）文書データに固有のコンテンツＩＤ、（２）タイトル、（３）本文などが関連付けられる。

【００１８】

固有表現・関係の学習データ編集部１３は、ユーザからの入力装置に対する操作に従い、文書ＤＢ１２に格納される学習用文書データ中の、抽出させる（抽出されるべき）固有表現に対応する文字列と、固有表現同士の関係として抽出させる固有表現の組とを指定（付与）することで、固有表現、固有表現同士の関係の学習データ（オリジナル学習文書）を生成する。この学習データは、オリジナル学習文書ＤＢ１４に格納される。

固有表現・関係の学習データ編集部１３は、固有表現として抽出させる文字列と、固有表現同士の関係として抽出させる固有表現の組とを指定する指定手段と呼ぶこともできる。

オリジナル学習文書ＤＢ１４に格納される学習データは、固有表現に係る学習データと、固有表現同士の関係に係る学習データとに区分される。

【００１９】

図３は、実施形態に係る固有表現抽出装置１０のオリジナル学習文書ＤＢ１４に格納される、固有表現に係る学習データの一例を表形式で示す図である。

図３に示した例では、オリジナル学習文書ＤＢ１４に格納される、固有表現に係る学習データは、（１）固有表現に固有のタグＩＤ、（２）固有表現が記述される文書データのコンテンツＩＤ、（３）タグの種類、（４）タグの値などが関連付けられる。

タグの種類は、固有表現の分類名、例えば「人名」、「地名」などである。タグの値は、具体的な固有表現の記述、例えば具体的な人名、地名などである。

【００２０】

図４は、実施形態に係る固有表現抽出装置１０のオリジナル学習文書ＤＢ１４に格納される、固有表現同士の関係に係る学習データの一例を表形式で示す図である。

図４に示した例では、オリジナル学習文書ＤＢ１４に格納される、固有表現同士の関係に係る学習データは、（１）固有表現同士の関係に固有の関係ＩＤ、（２）第１のタグＩＤ、（３）第１のタグの役割、（４）第２のタグＩＤ、（５）第２のタグの役割などが関

連付けられる。

【００２１】

図４に示した例では、関係ＩＤ「１」に関する関係として、タグＩＤが「１」である固有表現の役割「住人」と、タグＩＤが「２」である固有表現の役割「住んでいる地域」との間に関係が存在することが示される。また、この例では、関係ＩＤ「２」に関する関係として、タグＩＤが「３」である固有表現の役割「スポーツ」と、タグＩＤが「４」である固有表現の役割「順位」との間に関係が存在することが示される。

図４では、２種類の固有表現同士に存在する関係について定義された例について示したが、これに限らず３種類以上の固有表現同士に存在する関係について定義されてもよい。

【００２２】

学習データへの関係抽出ルール登録部１５は、抽出用文書データから関係が抽出されるべき固有表現の分類名（種類）の組を定める関係抽出ルールを、ＵＩに対するユーザからの入力操作に従って指定（登録）して、関係抽出ルールＤＢ１６に格納する。

【００２３】

図５は、実施形態に係る固有表現抽出装置１０の関係抽出ルールＤＢ１６に格納される関係抽出ルールの一例を表形式で示す図である。

図５に示した例では、関係抽出ルールＤＢ１６に格納される関係抽出ルールは、（１）関係抽出ルールに固有のルールＩＤ、（２）第１のタグの種類、（３）第２のタグの種類、（４）第１のタグの役割、（５）第２のタグの役割などが関連付けられる。関係抽出ルールＤＢ１６が設けられる不揮発性メモリは、関係抽出ルールを格納する格納手段と呼ぶこともできる。

【００２４】

図５に示した例では、ルールＩＤ「１」に関する関係として、第１のタグの種類「人名」と、第２のタグの種類「地名」と、第１のタグの役割「住人」と、第２のタグの役割「住んでいる地域」との間に関係が存在することが示される。また、この例では、ルールＩＤ「２」に関する関係として、第１のタグの種類「スポーツ」と、第２のタグの種類「順位」と、第１のタグの役割「競技名」と、第２のタグの役割「競技結果」との間に関係が存在することが示される。

【００２５】

学習データへの関係抽出ルール適用部１７は、オリジナル学習文書ＤＢ１４に格納される学習データに、関係抽出ルールＤＢ１６に格納される関係抽出ルールを適用することで、当該学習データにおける固有表現同士の関係のうち関係抽出ルールで定められる分類名で示される分類に属する固有表現同士の関係を一括で登録する。これにより、学習データへの関係抽出ルール適用部１７は、固有表現同士の関係が登録された学習データである学習文書を生成する。この学習文書は学習文書ＤＢ１８に格納される。学習データへの関係抽出ルール適用部１７は、学習文書を生成する生成手段と呼ぶこともできる。

学習文書ＤＢ１８に格納される学習文書の各項目は、上記のオリジナル学習文書ＤＢ１４に格納される各種の学習データ（図３、４参照）と同じである。

【００２６】

固有表現・関係抽出の学習部１９は、学習文書ＤＢ１８に格納された学習文書の内容を解析辞書ＤＢ２０に格納される解析辞書に反映することで、固有表現と固有表現同士の関係との抽出ルールを定めた抽出用辞書を学習する。

【００２７】

図６は、実施形態に係る固有表現抽出装置１０の解析辞書ＤＢ２０に格納される解析辞書の一例を表形式で示す図である。

図６に示した例では、解析辞書ＤＢ２０に格納される解析辞書（抽出用辞書）は、各行に固有の辞書ＩＤ、タグの種類、タグの特徴、タグの値、複数種類のタグ同士の関係などが関連付けられる。タグの特徴とは、タグの記述形式、例えばバイナリデータを示す。解析辞書ＤＢ２０が設けられる不揮発性メモリは、解析辞書を格納する格納手段と呼ぶこともできる。

【００２８】
この解析辞書は、抽出用文書データから固有表現と固有表現同士の関係とを抽出するために照合される辞書である。この解析辞書は、過去の学習用文書データに基づく学習結果が反映されており、また、新たな学習用文書データに基づく学習結果が反映される。この解析辞書は、ニューラルネットワークから構成される学習器であってもよい。
【００２９】
固有表現・関係抽出部２１は、解析辞書ＤＢ２０に格納される解析辞書と、文書ＤＢ１２に格納される抽出用文書データとを照合することで、抽出用文書データから固有表現と固有表現同士の関係をそれぞれ抽出する。
【００３０】
固有表現・関係抽出結果の表示部２２は、固有表現・関係抽出部２１による固有表現および固有表現同士の関係の抽出結果を表示装置に表示する。固有表現・関係抽出結果の表示部２２は、固有表現および固有表現同士の関係の抽出結果を出力する出力手段と呼ぶこともできる。
また、固有表現・関係抽出結果の表示部２２は、固有表現の抽出結果と固有表現同士の関係の抽出結果とを重ねて表示装置に表示することもできる。これにより、ユーザは、固有表現の誤抽出および未検出を発見しやすくなる。
【００３１】
（第１の処理）
次に、固有表現抽出装置１０の第１の処理について説明する。
図７は、実施形態に係る固有表現抽出装置１０の第１の処理手順の一例を示すフローチャートである。
まず、ユーザからの入力操作に従って、文書受付部１１は、学習用文書データの登録を受け付けて、登録した学習用文書データを文書ＤＢ１２に格納する（Ｓ１１）。
【００３２】
文書ＤＢ１２に格納された学習用文書データは表示装置に表示される。この表示された状態で、表示画面上の学習用文書データの本文の記述に対するユーザからの入力操作に従って、固有表現・関係の学習データ編集部１３は、学習用文書データにおける、ユーザの入力操作により指定された記述にタグであることを示すマーク（下線）を付与する（以下、タグを付与すると称することがある）。タグの付与により生成された、固有表現に係る学習データ（図３参照）はオリジナル学習文書ＤＢ１４に格納される（Ｓ１２）。なお、固有表現同士の関係に係る学習データ（図４参照）は、第１の処理では生成されない。
【００３３】
図８は、実施形態に係る固有表現抽出装置１０による、文書データの記述へのタグの付与時の表示画面Ｇ１の一例を示す図である。
図８に示した例では、固有表現抽出装置１０の表示装置に表示された画面Ｇ１上の文書データの本文中の各記述「搬送異常」、「ボルト」、「ボルトが緩んでいます。」、「アームに付いているネジを締めました。」への画面上のポインタによる指定などにより、各記述にタグをそれぞれ付与することができる。
【００３４】
画面Ｇ１上の分類名にかかるウインドウに対するポインタによる指定などにより、タグが付与される各記述には当該タグの分類名をあわせて付与できる。図８に示した例では、タグが付与された上記の記述「搬送異常」にはタグの分類名「現象」を、上記の記述「ボルト」には分類名「部位」を、上記の記述「ボルトが緩んでいます。」には分類名「原因」を、上記の記述「アームに付いているネジを締めました。」には分類名「対処」をそれぞれ付与できる。
【００３５】
また、画面Ｇ１と異なる図示しない設定画面上でのユーザからの入力操作に従って、学習データへの関係抽出ルール登録部１５は、タグの任意の第１の分類名とタグの任意の第２の分類名との間に関係（リンク）を付与する。この付与により生成される関係抽出ルー

ル（図５参照）は関係抽出ルールＤＢ１６に格納される（Ｓ１３）。この設定画面は、画面Ｇ１における学習用文書データの表示と並べて表示することができる。第１の処理におけるタグの任意の分類名同士の関係の付与は、上記の学習用文書データの記述によらない付与である。この付与は、１つの分類名と複数の分類名との間で行なうこともできる。

【００３６】

ここで、学習データへの関係抽出ルール適用部１７は、以下条件の時に、関係抽出ルールＤＢ１６に格納される関係抽出ルールの登録内容に合わせて、オリジナル学習文書ＤＢ１４に格納される学習データで示されるタグのうち、後述のある分類名に係るタグと、別の分類名に係るタグとの関係を追加、編集、または削除する処理を行なう。上述の条件とは、Ｓ１３において、（１）学習データへの関係抽出ルール登録部１５によって、関係抽出ルールに対して、ある分類名に係るタグと、別の分類名に係るタグとの間の関係付与（登録）が完了している場合（Ｓ１４のＮＯ）、または、（２）完了する前で、当該付与を新たに行なう場合（Ｓ１５のＹＥＳ）である。この処理により生成された学習文書は学習文書ＤＢ１８に格納される（Ｓ１６）。例えば、関係抽出ルールで、分類名Ａと分類名Ｂとの関係が定義されていれば、学習データにおける、分類名Ａに属するタグと分類名Ｂに属するタグとの間に関係が付与されることになる。

上記の第１の処理によれば、学習データで示される、ある分類名に係るタグと、別の分類名に係るタグとの関係を一括で登録できる。

【００３７】

（第２の処理）

次に、固有表現抽出装置１０の第２の処理について説明する。

図９は、実施形態に係る固有表現抽出装置１０の第２の処理手順の一例を示すフローチャートである。

第２の処理では、まず、固有表現・関係抽出結果の表示部２２は、固有表現・関係抽出部２１により抽出用文書データから抽出したタグ、およびタグ同士の関係を分類名ごとにグループ化した抽出結果の表示画面Ｇ２を表示装置に表示する（Ｓ２１）。固有表現・関係抽出部２１による抽出結果と抽出元文書データとの関係を示す情報は、固有表現・関係抽出結果の表示部２２に接続される内部メモリに格納されているとする。

【００３８】

図１０は、実施形態に係る固有表現抽出装置１０による、抽出されたタグ、およびタグ同士の関係の表示画面Ｇ２の一例を示す図である。

図１０に示した表示画面Ｇ２では、分類（分類名）Ａ，Ｂ，Ｃ，Ｄなどに属する複数種類のタグが示され、ある分類に属するタグと異なる分類に属するタグとの間の関係を示す。

図１０では、分類Ａと分類Ｂとの間、分類Ｂと分類Ｃとの間、分類Ｃと分類Ｄとの間でのタグ同士の関係がそれぞれ設定される例を示すが、これに限らず、例えば分類Ａと分類Ｃとの間、分類Ｂと分類Ｄとの間などでのタグ同士の関係が設定されてもよい。

【００３９】

ユーザは、抽出結果の表示画面Ｇ２で示される、気になるタグ、またはタグ同士との関係を入力操作により指定することができる（Ｓ２２）。気になるタグ、またはタグ同士の関係とは、抽出用文書データからの抽出結果として適切でない可能性があるタグ、またはタグ同士の関係である。Ｓ２２での指定に伴い、固有表現・関係抽出結果の表示部２２は、上記の内部メモリに格納される、固有表現・関係抽出部２１による抽出結果と抽出元文書データとの関係を示す情報を固有表現・関係の学習データ編集部１３に渡す。

【００４０】

Ｓ２２での指定を受けて、固有表現・関係の学習データ編集部１３は、指定されたタグ、またはタグ同士の関係の抽出元文書データを上記の渡された情報から検索して、この検索された抽出元文書データの本文などを表示装置に表示する（Ｓ２３）。

【００４１】

この表示を受けて、ユーザからの入力操作により、固有表現・関係の学習データ編集部

１３は、抽出元文書データの記述に付与されているタグ、またはタグとタグの関係を編集する（Ｓ２４）。

【００４２】

図１１は、実施形態に係る固有表現抽出装置１０による、抽出されたタグ、およびタグ同士の関係の編集画面Ｇ３の一例を示す図である。

図１１に示した例では、分類Ｂに属する１つ目のタグと分類Ｃに属する１つ目のタグ同士の関係が編集対象として指定された例を示す。この画面Ｇ３では、ユーザからの入力操作に従って、固有表現・関係の学習データ編集部１３は、指定された関係の変更、例えば分類Ｂに属する１つ目のタグと、分類Ｃに属する２つ目以降のタグ同士の関係への修正、または関係の削除などを行なうことができる。

また、上記のように、付与済みのタグ自体の修正または削除などを行なうこともできる。タグ自体の修正とは、例えば分類名の修正、対象となる記述の変更である。タグ自体の削除とは、対象となる記述に対する固有表現としての指定の解除である。

第２の処理によれば、タグ、およびタグとタグとの関係の抽出結果のうち、指定された抽出結果の抽出元文書を容易に表示することができる。また、タグとタグとの関係の確認、編集を容易に行うことができる。

【００４３】

（第３の処理）

次に、固有表現抽出装置１０の第３の処理について説明する。

図１２は、実施形態に係る固有表現抽出装置１０の第３の処理手順の一例を示すフローチャートである。

まず、ユーザからの入力操作に従って、文書受付部１１は、学習用文書データの登録を受け付けて、登録した学習用文書データを文書ＤＢ１２に格納する（Ｓ３１）。ここでは抽出用文書データは文書ＤＢ１２に格納済みであるとする。

【００４４】

文書ＤＢ１２に格納される学習用文書データは表示装置に表示される。この表示された状態で、表示画面上の学習用文書データの記述に対するユーザからの入力操作に従って、固有表現・関係の学習データ編集部１３は、学習用文書データの記述にタグを付与する。タグの付与により生成された、固有表現に係る学習データ（図３参照）はオリジナル学習文書ＤＢ１４に格納される。

ここでは、学習用文書データの記述へのタグの付与時の表示画面は図８に示した表示画面Ｇ１であるとする。

【００４５】

この表示画面Ｇ１に表示される学習用文書データの記述に対するユーザからの入力操作にしたがって、固有表現・関係の学習データ編集部１３は、学習用文書データにおける記述に付与された第１のタグと第２のタグとの間に関係（リンク）を付与する。この付与により生成された、固有表現同士の関係に係る学習データ（図４参照）はオリジナル学習文書ＤＢ１４に格納される（Ｓ３２）。第３の処理におけるタグ同士の関係の付与は、上記の学習用文書データの本文の記述に対する付与である。ここでの関係の付与は、１つのタグと複数のタグとの間で行なうこともできる。

【００４６】

図１３は、実施形態に係る固有表現抽出装置１０による、文書データのタグ同士の関係の付与時の表示画面Ｇ４の一例を示す図である。

図１３では、表示装置に表示された画面Ｇ４上の文書データの本文中の第１の記述「ボルトが緩んでいます。」に付与された、分類名「原因」に係るタグと、本文中の第２の記述「アームに付いているネジを締めました。」に付与された、分類名「対処」に係るタグとの関係を示す線Ｌ１が付与された例を示す。

【００４７】

第３の処理では、第１の処理で説明した、学習データへの関係抽出ルール登録部１５による処理は行なわれず、Ｓ３２でオリジナル学習文書ＤＢ１４に格納された各種学習デー

タは、学習データへの関係抽出ルール適用部１７を介して学習文書として学習文書ＤＢ１８に格納される。

【００４８】

次に、固有表現・関係抽出の学習部１９は、学習文書ＤＢ１８に格納された学習文書の内容を解析辞書ＤＢ２０に格納される解析辞書に反映することで、固有表現と固有表現同士の関係との抽出ルールを学習する（Ｓ３３）。

【００４９】

固有表現・関係抽出部２１は、解析辞書ＤＢ２０に格納される解析辞書を用いて、文書ＤＢ１２に格納される抽出用文書データから、タグ、およびタグとタグの関係をそれぞれ抽出する（Ｓ３４）。

【００５０】

固有表現・関係抽出結果の表示部２２は、Ｓ３４で抽出されたタグ、およびタグ同士の関係を分類名ごとにグループ化した抽出結果の表示画面Ｇ２を表示装置に表示する（Ｓ３５）。

【００５１】

固有表現・関係抽出結果の表示部２２は、学習文書ＤＢ１８に格納された学習文書とＳ３４での抽出結果とを照合する。この照合により、固有表現・関係抽出結果の表示部２２は、固有表現・関係抽出の学習部１９により学習文書として生成されたが、Ｓ３４で当該抽出用文書データから抽出されなかったタグ、およびタグとタグの関係を特定し、この特定した結果を示す表示画面Ｇ５を表示装置に表示する（Ｓ３６）。

上記の、学習文書として生成されたが、抽出用文書データから抽出されなかったタグ、およびタグとタグの関係は、例えば、固有表現・関係抽出の学習部１９による解析辞書への学習の不具合、ここでは解析辞書に反映させる定義の欠落などに起因して生ずる。

【００５２】

図１４は、実施形態に係る固有表現抽出装置１０による、抽出されなかったタグ、およびタグ同士の関係の表示画面Ｇ５の一例を示す図である。

図１４に示した例では、点線で囲まれる、分類Ｂに属する１つ目のタグ、分類Ｃに属する１つ目および３つ目のタグは、学習文書に含まれていたが抽出用文書データから抽出されなかったタグとして示される。

【００５３】

また、図１４に示した例では、点線で示される、（１）分類Ａに属する２つ目のタグと分類Ｂに属する１つ目のタグとの間の関係、（２）分類Ｂに属する３つ目のタグと分類Ｃに属する３つ目のタグとの間の関係、および（３）分類Ｃに属する２つ目のタグと分類Ｄに属する３つ目のタグとの間の関係は、学習文書に含まれていたが抽出用文書データから抽出されなかった関係として示される。

第３の処理により、タグ、およびタグとタグの関係の抽出結果の抽出漏れを容易に確認することができる。

【００５４】

（第４の処理）

次に、固有表現抽出装置１０の第４の処理について説明する。

図１５は、実施形態に係る固有表現抽出装置１０の第４の処理手順の一例を示すフローチャートである。

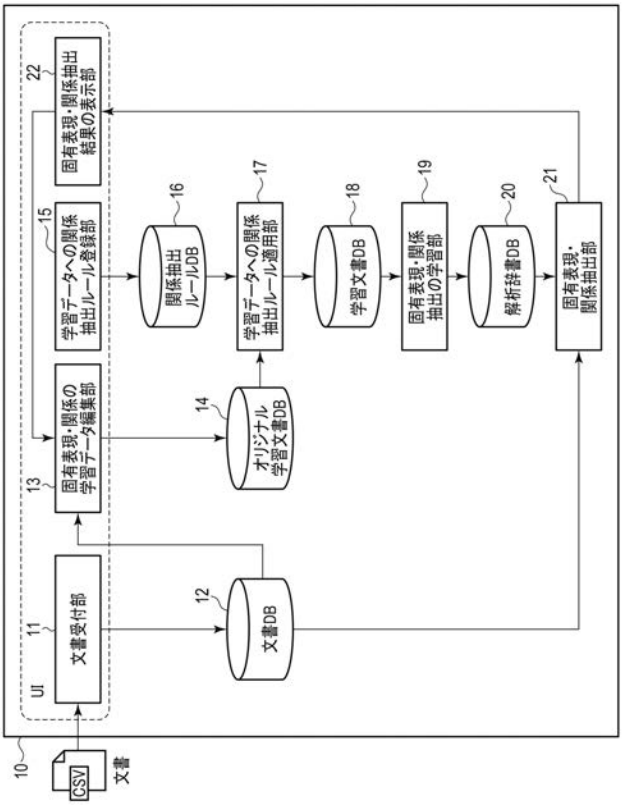
第４の処理では、第３の処理で説明したＳ３１～Ｓ３５までの処理がなされる（Ｓ４１～Ｓ４５）。

【００５５】

そして、固有表現・関係抽出結果の表示部２２は、学習文書ＤＢ１８に格納された学習文書とＳ４４（Ｓ３４と同様）での抽出結果とを照合する。この照合により、固有表現・関係抽出結果の表示部２２は、固有表現・関係抽出の学習部１９により学習文書として生成されておらず、直近の学習された解析辞書にも定義されていないが、Ｓ４４で当該抽出用文書データから抽出されたタグ、およびタグとタグの関係を特定し、この特定した結果

の表示画面Ｇ６を表示装置に表示する（Ｓ４６）。
上記の、学習文書として生成されなかったが、抽出用文書データから抽出されたタグ、およびタグとタグの関係は、例えば、固有表現・関係抽出の学習部１９による解析辞書への学習の不具合、ここでは解析辞書に対する不必要な定義の追加などに起因して生ずる。
【００５６】
図１６は、実施形態に係る固有表現抽出装置１０による、学習されなかったが抽出されたタグ、およびタグ同士の関係の表示画面Ｇ６の一例を示す図である。
図１６に示した例では、二重線で囲まれる、分類Ｃに属する１つ目のタグは、学習文書には含まれなかったが抽出用文書データから抽出されたタグとして示される。
【００５７】
また、図１６に示した例では、二重線で示される、（１）分類Ａに属する２つ目のタグと分類Ｂに属する１つ目のタグとの間の関係、（２）分類Ｂに属する１つ目のタグと分類Ｃに属する１つ目のタグとの間の関係が示される。これらの関係は、学習文書には含まれなかったが抽出用文書データから抽出された関係として示される。
第４の処理により、タグ、およびタグとタグの関係の抽出結果の誤抽出を容易に確認することができる。
【００５８】
以上説明したように、実施形態に係る固有表現抽出装置は、学習データにおけるタグ同士の関係を一括で登録したり、抽出元文書を容易に表示したり、抽出結果の抽出漏れ又は誤抽出を容易に確認できたりするので、文書からの固有表現抽出の精度を向上させることができる。
【００５９】
本発明のいくつかの実施形態を説明したが、これらの実施形態は、例として提示したものであり、発明の範囲を限定することは意図していない。これら新規な実施形態は、その他の様々な形態で実施されることが可能であり、発明の要旨を逸脱しない範囲で、種々の省略、置き換え、変更を行うことができる。これら実施形態やその変形は、発明の範囲や要旨に含まれるとともに、特許請求の範囲に記載された発明とその均等の範囲に含まれる。
【００６０】
また、各実施形態に記載した手法は、計算機（コンピュータ）に実行させることができるプログラム（ソフトウェア手段）として、例えば磁気ディスク（フロッピー（登録商標）ディスク、ハードディスク等）、光ディスク（ＣＤ－ＲＯＭ、ＤＶＤ、ＭＯ等）、半導体メモリ（ＲＯＭ、ＲＡＭ、フラッシュメモリ等）等の記録媒体に格納し、また通信媒体により伝送して頒布することもできる。なお、媒体側に格納されるプログラムには、計算機に実行させるソフトウェア手段（実行プログラムのみならずテーブルやデータ構造も含む）を計算機内に構成させる設定プログラムをも含む。本装置を実現する計算機は、記録媒体に記録されたプログラムを読み込み、また場合により設定プログラムによりソフトウェア手段を構築し、このソフトウェア手段によって動作が制御されることにより上述した処理を実行する。なお、本明細書でいう記録媒体は、頒布用に限らず、計算機内部あるいはネットワークを介して接続される機器に設けられた磁気ディスクや半導体メモリ等の記憶媒体を含むものである。
【符号の説明】
【００６１】
１０…固有表現抽出装置、１１…文書受付部、１２…文書ＤＢ、１３…固有表現・関係の学習データ編集部、１４…オリジナル学習文書ＤＢ、１５…学習データへの関係抽出ルール登録部、１６…関係抽出ルールＤＢ、１７…学習データへの関係抽出ルール適用部、１８…学習文書ＤＢ、１９…固有表現・関係抽出の学習部、２０…解析辞書ＤＢ、２１…固有表現・関係抽出部、２２…固有表現・関係抽出結果の表示部。

【図１】



【図２】

(文書DB)

コンテンツID	タイトル	本文	…
1	…	…	…
2	…	…	…
…	…	…	…

【図３】

(オリジナル学習文書DB(固有表現))

タグID	コンテンツID	タグの種類	タグの値	…
1	1	人名	○芝太郎	…
2	1	地名	○芝町	…
…	…	…	…	…

【図４】

(オリジナル学習文書DB(関係))

関係ID	第1のタグID	第1のタグの役割	第2のタグID	第2のタグの役割	…
1	1	住人	2	住んでいる地域	…
2	3	スポーツ	4	順位	…
…	…	…	…	…	…

【図５】

(関係抽出ルールDB(固有表現))

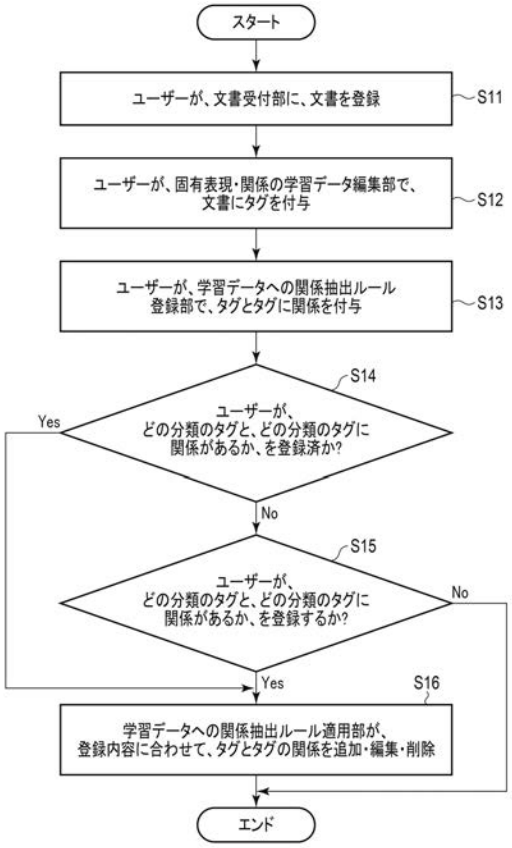
ルールID	第1のタグの種類	第2のタグの種類	第1のタグの役割	第2のタグの役割	…
1	人名	地名	住人	住んでいる地域	…
2	スポーツ	順位	競技名	競技結果	…
…	…	…	…	…	…

【図６】

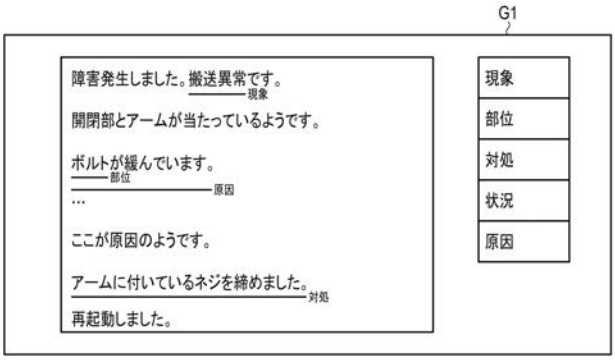
(解析辞書DB)

辞書ID	タグの種類	タグの特徴	…
1	人名	バイナリデータ	…
2	地名	バイナリデータ	…
…	…	…	…

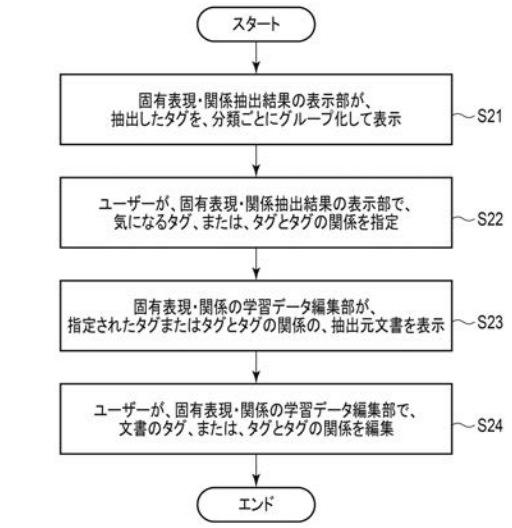
【図７】


スタート
ユーザーが、文書受付部に、文書を登録
S11
ユーザーが、固有表現・関係の学習データ編集部で、
文書にタグを付与
S12
ユーザーが、学習データへの関係抽出ルール
登録部で、タグとタグに関係を付与
S13
S14
ユーザーが、
どの分類のタグと、どの分類のタグに
関係があるか、を登録済か?
Yes
No
S15
ユーザーが、
どの分類のタグと、どの分類のタグに
関係があるか、を登録するか?
No
Yes
S16
学習データへの関係抽出ルール適用部が、
登録内容に合わせて、タグとタグの関係を追加・編集・削除
エンド

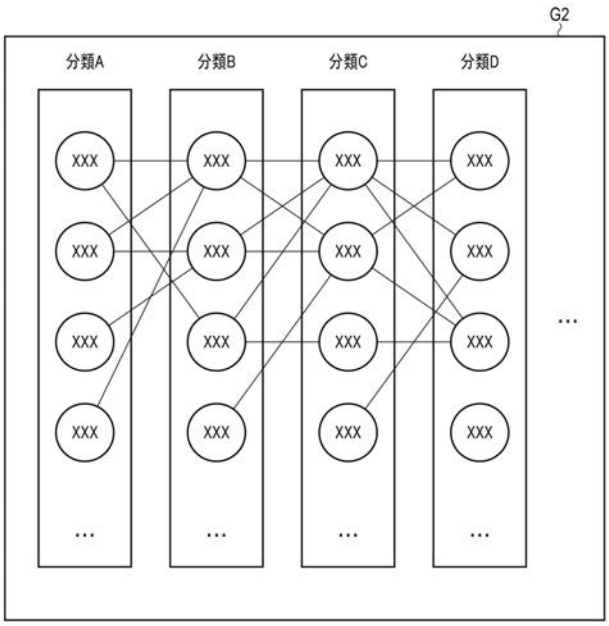
【図８】


G1
障害発生しました。搬送異常です。
現象
開閉部とアームが当たっているようです。
ボルトが緩んでいます。
部位
原因
…
ここが原因のようです。
アームに付いているネジを締めました。
対処
再起動しました。
現象
部位
対処
状況
原因

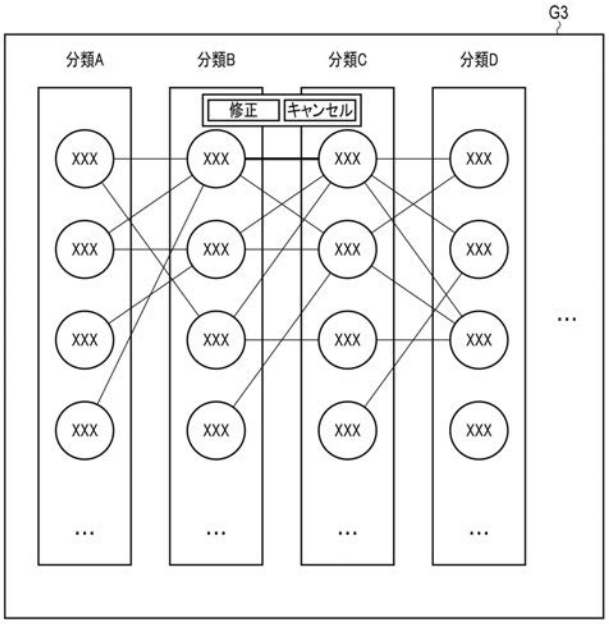
【図９】


スタート
固有表現・関係抽出結果の表示部が、
抽出したタグを、分類ごとにグループ化して表示
S21
ユーザーが、固有表現・関係抽出結果の表示部で、
気になるタグ、または、タグとタグの関係を指定
S22
固有表現・関係の学習データ編集部が、
指定されたタグまたはタグとタグの関係の、抽出元文書を表示
S23
ユーザーが、固有表現・関係の学習データ編集部で、
文書のタグ、または、タグとタグの関係を編集
S24
エンド

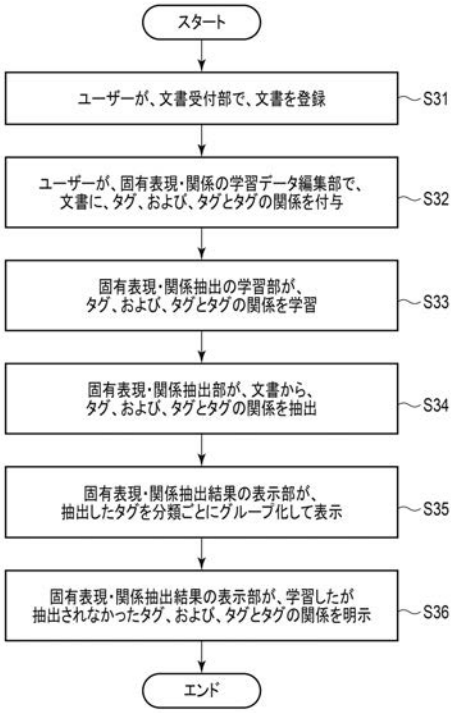
【図１０】


G2
分類A
分類B
分類C
分類D
XXX
…

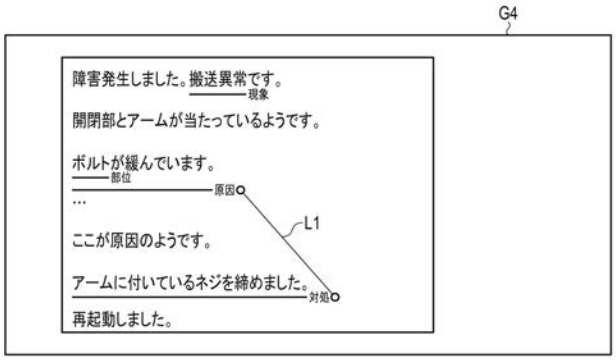
【図１１】



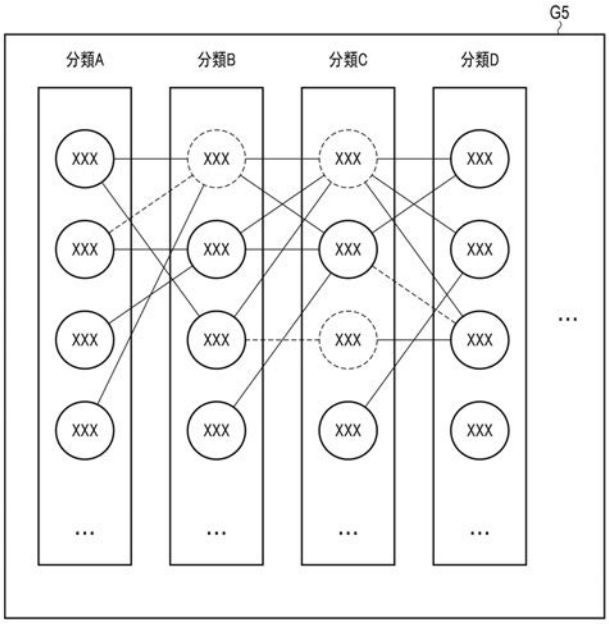
【図１２】



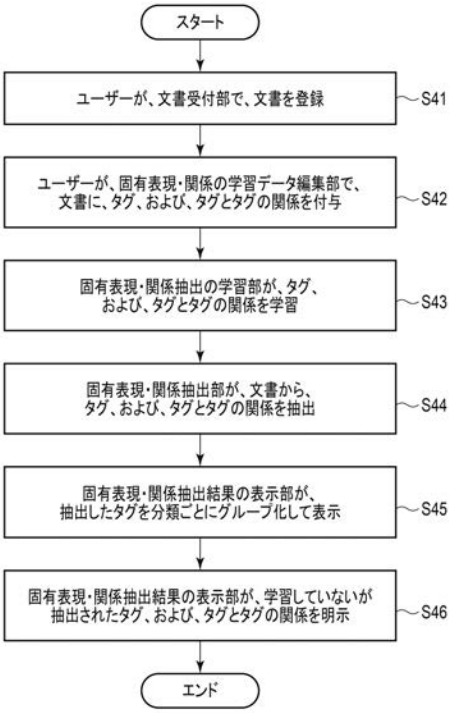
【図１３】



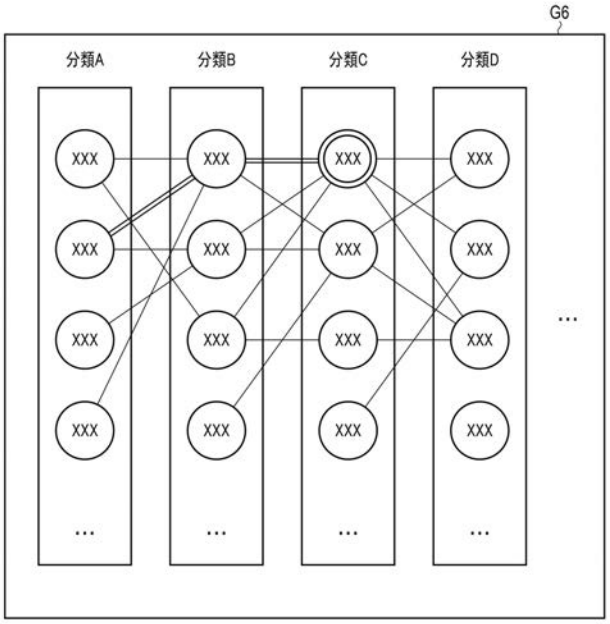
【図１４】



【図１５】


スタート
ユーザーが、文書受付部で、文書を登録
S41
ユーザーが、固有表現・関係の学習データ編集部で、
文書に、タグ、および、タグとタグの関係を付与
S42
固有表現・関係抽出の学習部が、タグ、
および、タグとタグの関係を学習
S43
固有表現・関係抽出部が、文書から、
タグ、および、タグとタグの関係を抽出
S44
固有表現・関係抽出結果の表示部が、
抽出したタグを分類ごとにグループ化して表示
S45
固有表現・関係抽出結果の表示部が、学習していないが
抽出されたタグ、および、タグとタグの関係を明示
S46
エンド

【図１６】


G6
分類A
分類B
分類C
分類D
XXX
…

フロントページの続き

(74)代理人　100189913
弁理士　鵜飼　健
(72)発明者　飛田　義賢
神奈川県川崎市幸区堀川町７２番地３４　東芝デジタルソリューションズ株式会社内
(72)発明者　鈴木　優
神奈川県川崎市幸区堀川町７２番地３４　東芝デジタルソリューションズ株式会社内
Ｆターム(参考) 5B091 AA15　AB06　CA01　CC03　EA01