

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4940325号  
(P4940325)

(45) 発行日 平成24年5月30日(2012.5.30)

(24) 登録日 平成24年3月2日(2012.3.2)

(51) Int.Cl. F I  
**G06F 17/21 (2006.01)** G O 6 F 17/21 5 5 O A  
**G06F 17/27 (2006.01)** G O 6 F 17/27 Z

請求項の数 11 (全 14 頁)

(21) 出願番号	特願2010-76047 (P2010-76047)	(73) 特許権者	000003078
(22) 出願日	平成22年3月29日 (2010.3.29)		株式会社東芝
(65) 公開番号	特開2011-209941 (P2011-209941A)		東京都港区芝浦一丁目1番1号
(43) 公開日	平成23年10月20日 (2011.10.20)	(74) 代理人	100108855
審査請求日	平成23年9月16日 (2011.9.16)		弁理士 蔵田 昌俊
		(74) 代理人	100091351
			弁理士 河野 哲
		(74) 代理人	100088683
			弁理士 中村 誠
		(74) 代理人	100109830
			弁理士 福原 淑弘
		(74) 代理人	100075672
			弁理士 峰 隆司
		(74) 代理人	100095441
			弁理士 白根 俊郎

最終頁に続く

(54) 【発明の名称】 文書校正支援装置、方法およびプログラム

(57) 【特許請求の範囲】

【請求項1】

校正処理の対象となる1以上の文を含む校正文書または登録処理の対象となる1以上の文を含む登録文書の入力を受け付ける文書入力部と、

前記登録文書に含まれる文を形態素解析、構文解析、および係り受け解析して単語間の意味的なつながりを示す依存関係を含む解析情報を生成し、前記校正文書に含まれる文を形態素解析して形態素解析情報を生成する文書解析部と、

前記形態素解析情報を参照して、名詞となりうる2以上の語が連続する関係である名詞連続関係を有する複合語を造語候補文字列として検出する造語候補検出部と、

前記登録文書に含まれる文と前記解析情報とを対応付けた構文情報を格納する構文情報データベースと、

前記構文情報を参照して、前記造語候補文字列を構成する構成語が格要素となり、かつ該構成語が前記名詞連続関係以外の前記依存関係を有する依存関係文があるかどうかを検索し、該依存関係文があれば前記造語候補文字列を造語と判定する構文情報検索部と、

前記造語を含むメッセージを表示する警告表示部と、を具備することを特徴とする文書校正支援装置。

【請求項2】

前記造語候補検出部は、前記校正文書が日本語である場合、名詞またはサ変名詞の2以上の語が連続する複合語を造語候補文字列として検出することを特徴とする請求項1に記載の文書校正支援装置。

## 【請求項 3】

前記造語候補検出部は、前記校正文書が中国語である場合、名詞、または名詞および動詞の両方の解釈がある語の 2 以上の語が連続する複合語を造語候補文字列として検出することを特徴とする請求項 1 に記載の文書校正支援装置。

## 【請求項 4】

前記構文情報検索部は、前記校正文書が日本語である場合、1 以上の前記構成語が助詞の格要素となり、かつ該構成語間で前記名詞連続関係以外の依存関係を有する前記依存関係文があるかどうかを検索することを特徴とする請求項 1 または請求項 2 に記載の文書校正支援装置。

## 【請求項 5】

前記構文情報検索部は、前記校正文書が中国語である場合、1 以上の前記構成語が動詞の格要素となり、かつ該構成語間で前記名詞連続関係以外の依存関係を有する前記依存関係文があるかどうかを検索することを特徴とする請求項 1 または請求項 3 に記載の文書校正支援装置。

## 【請求項 6】

前記警告表示部は、前記造語と前記依存関係文とを含む警告メッセージを表示することを特徴とする請求項 1 から請求項 5 のいずれか 1 項に記載の文書校正支援装置。

## 【請求項 7】

前記構文情報にさらに、前記登録文書に含まれる文が過去に校正されたかどうかを示す校正情報を含めて前記構文情報データベースに格納する構文情報管理部をさらに具備することを特徴とする請求項 1 から請求項 6 のいずれか 1 項に記載の文書校正支援装置。

## 【請求項 8】

前記造語候補検出部は、前記依存関係文の校正情報が校正済であることを示す場合は、前記造語候補文字列を造語として検出しないことを特徴とする請求項 7 に記載の文書校正支援装置。

## 【請求項 9】

前記構文情報データベースは、格と見出し語とで構成された木構造を格納し、

前記構文情報検索部は、前記木構造を参照することにより前記構成語間の依存関係を検索することを特徴とする請求項 1 から請求項 8 に記載の文書校正支援装置。

## 【請求項 10】

校正処理の対象となる 1 以上の文を含む校正文書または登録処理の対象となる 1 以上の文を含む登録文書の入力を受け付け、

前記登録文書に含まれる文を形態素解析、構文解析、および係り受け解析して単語間の意味的なつながりを示す依存関係を含む解析情報を生成し、前記校正文書に含まれる文を形態素解析して形態素解析情報を生成し、

前記形態素解析情報を参照して、名詞となりうる 2 以上の語が連続する関係である名詞連続関係を有する複合語を造語候補文字列として検出し、

前記登録文書に含まれる文と前記解析情報とを対応付けた構文情報を格納する構文情報データベースを用意し、

前記構文情報を参照して、前記造語候補文字列を構成する構成語が格要素となり、かつ該構成語が前記名詞連続関係以外の前記依存関係を有する依存関係文があるかどうかを検索し、該依存関係文があれば前記造語候補文字列を造語と判定し、

前記造語を含むメッセージを表示することを具備することを特徴とする文書校正支援方法。

## 【請求項 11】

コンピュータを、

校正処理の対象となる 1 以上の文を含む校正文書または登録処理の対象となる 1 以上の文を含む登録文書の入力を受け付ける文書入力手段と、

前記登録文書に含まれる文を形態素解析、構文解析、および係り受け解析して単語間の意味的なつながりを示す依存関係を含む解析情報を生成し、前記校正文書に含まれる文を

10

20

30

40

50

形態素解析して形態素解析情報を生成する文書解析手段と、

前記形態素解析情報を参照して、名詞となりうる2以上の語が連続する関係である名詞連続関係を有する複合語を造語候補文字列として検出する造語候補検出手段と、

前記登録文書に含まれる文と前記解析情報とを対応付けた構文情報をデータベースに格納する格納手段と、

前記構文情報を参照して、前記造語候補文字列を構成する構成語が格要素となり、かつ該構成語が前記名詞連続関係以外の前記依存関係を有する依存関係文があるかどうかを検索し、該依存関係文があれば前記造語候補文字列を造語と判定する構文情報検索手段と、

前記造語を含むメッセージを表示する警告表示手段として機能させるための文書校正支援プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、技術文書の校正支援システムに関し、特に造語の検出機能を有する文書校正支援装置、方法およびプログラムに関する。

【背景技術】

【0002】

技術者が報告書や論文、特許などの技術文書を作成し、この技術文書はその分野の専門家でない翻訳者が他言語に翻訳する場合がある。文書作成者が使用した複合語(造語)を翻訳する際、文書を作成した専門家には意味または内容が明らかな用語でも、翻訳者にとっては前提となる技術的知識の欠如によって、どのように翻訳すればよいのかわからないといった問題がある。この場合、複合語やその部分単語列からなる複合語のコーパス頻度を調べ、その頻度が低いものを利用頻度が小さい複合語、すなわち造語として警告するという方法がある(例えば、特許文献1参照)。

【先行技術文献】

【特許文献】

【0003】

【特許文献1】特開2001-249921号公報

【発明の概要】

【発明が解決しようとする課題】

【0004】

しかしながら、コーパス頻度を用いるだけでは翻訳者の翻訳作業において訳語の判断が難しい場合がある。例えば、名詞「事象」とサ変名詞「抽出」とを組み合わせた造語「事象抽出」という複合語が使用された場合を仮定する。この場合、「事象抽出」という用語のコーパス頻度が予め設定した閾値よりも大きければ、この複合語は造語として判定されない。しかし、専門家でない翻訳者が「事象抽出」という用語を他言語、例えば英語に翻訳しようする場合、「事象」の訳語を“event”、「抽出」の訳語を“extraction”とすると、「事象抽出」を“event extraction”と訳せばよいのか、“extraction from event”と訳せばよいのか、あるいは“extraction of event”と訳せばよいのかという判断が難しい。

また、複合語「事象抽出」が造語かどうかを判定する方法として、例えば単純に「事象」「抽出」の2語を用いた文字列検索や「事象\*抽出」のようなワイルドカード検索が考えられる。しかし、これらの方法では、「事象データ系列抽出」または「事象データにより関連性が抽出される」などのような単語列が検索される可能性があり、「事象抽出」という複合語が造語かどうかの判定は難しい。よって、「事象抽出」という用語の使用頻度だけでは、翻訳者はその用語が造語であるかどうかを判断することができず、翻訳者から執筆者への問い合わせ、執筆者からの回答、および元文書の校正作業が必要となり、結果として翻訳終了までに多くの時間を要してしまう問題がある。

【0005】

本発明は、上述の課題を解決するためになされたものであり、文書校正の精度を高め作

10

20

30

40

50

業効率を向上させることができる文書校正支援装置、方法およびプログラムを提供することを目的とする。

【課題を解決するための手段】

【0006】

上述の課題を解決するため、本発明に係る文書校正支援装置は、校正処理の対象となる1以上の文を含む校正文書または登録処理の対象となる1以上の文を含む登録文書の入力を受け付ける文書入力部と、前記登録文書に含まれる文を形態素解析、構文解析、および係り受け解析して単語間の意味的なつながりを示す依存関係を含む解析情報を生成し、前記校正文書に含まれる文を形態素解析して形態素解析情報を生成する文書解析部と、前記形態素解析情報を参照して、名詞となりうる2以上の語が連続する関係である名詞連続関係  
10  
を有する複合語を造語候補文字列として検出する造語候補検出部と、前記登録文書に含まれる文と前記解析情報とを対応付けた構文情報を格納する構文情報データベースと、前記構文情報を参照して、前記造語候補文字列を構成する構成語が格要素となり、かつ該構成語が前記名詞連続関係以外の前記依存関係を有する依存関係文があるかどうかを検索し、該依存関係文があれば前記造語候補文字列を造語と判定する構文情報検索部と、前記造語を含むメッセージを表示する警告表示部と、を具備することを特徴とする。

【発明の効果】

【0007】

本発明の文書校正支援装置、方法およびプログラムによれば、文書校正の精度を高め作業効率を向上させることができる。  
20

【図面の簡単な説明】

【0008】

【図1】本実施形態に係る文書校正支援装置を示すブロック図。

【図2】本実施形態に係る文書校正支援装置の造語判定処理を示すフローチャート。

【図3】構文情報データベースに格納される構文情報の一例を示す図。

【図4】インデックステーブルの一例を示す図。

【図5】本実施形態の変形例に係る文書校正支援装置の造語判定処理を示すフローチャート。

【発明を実施するための形態】

【0009】

以下、図面を参照しながら本発明の実施形態に係る文書校正支援装置、方法およびプログラムについて詳細に説明する。なお、以下の実施形態では、同一の番号を付した部分については同様の動作を行うものとして、重ねての説明を省略する。  
30

次に、本実施形態に係る文書校正支援装置について図1を参照して詳細に説明する。

本実施形態に係る文書校正支援装置100は、文書入力部101、文書解析部102、造語候補検出部103、構文情報管理部104、構文情報データベース(DB)105、構文情報検索部106、校正制御部107、および警告表示部108を含む。

【0010】

文書入力部101は、文書の校正処理を行う場合に校正処理の対象となる校正文書の入力を受け付け、文書の登録処理を行う場合に登録処理の対象となる登録文書の入力を受け付ける。登録文書または校正文書として入力される文書は、文書全体またはその一部分から成る1以上の文を含む文書である。文書は、日本語による文書だけに限らずどのような言語で書かれていてもよく、例えば中国語、韓国語、フランス語で書かれた文書でもよい。  
40

【0011】

文書解析部102は、文書の登録処理を行う場合(すなわち、文書入力部101に登録文書が入力された場合)は、後述する校正制御部107から登録文書を受け取り、登録文書中の各文を形態素解析、構文解析、および係り受け解析して、登録文書中の各文に対する解析情報を生成する。解析情報には、単語間の意味的なつながりを示す依存関係が含まれる。依存関係の具体例としては、例えば「犬が歩く」という文には、「犬」と「歩く」  
50

との単語間にガ格によって結ばれる依存関係がある。また、文書解析部 102 は、文書の校正処理を行う場合（すなわち、文書入力部 101 に校正文書が入力された場合）は、校正制御部 107 から校正文書を受け取り、校正文書中の各文を形態素解析して形態素解析情報を生成する。なお、形態素解析、構文解析、および係り受け解析については、一般的な処理をそれぞれ用いるため、ここでの詳細な説明は省略する。

【0012】

造語候補検出部 103 は、文書の校正処理を行う場合は、文書解析部 102 から校正文書中の各文に対する形態素解析情報を受け取り、形態素解析情報を参照して造語候補文字列があるかどうかを検出する。造語候補文字列は、名詞となりうる 2 以上の語が連続する複合語（以下、名詞連続関係を有する複合語ともいう）を示す。名詞となりうる語としては、例えば名詞、サ変名詞がある。サ変名詞の具体例としては、「処理」という単語があり、「処理 + する」といった名詞に「する」をつけると動詞にもなる単語である。

10

【0013】

構文情報管理部 104 は、文書の登録処理を行う場合は、文書解析部 102 から後述する校正制御部 107 を通じて登録文書中の文とその解析情報とを受け取り、解析情報を後述する構文情報データベース 105 に登録するための管理を行う。

【0014】

構文情報データベース 105 は、文書の登録処理を行う場合は、構文情報管理部 104 から登録文書中の各文に対する解析情報を受け取り、登録文書中の文とその解析情報とを構文情報として格納する。構文情報データベース 105 に格納される構文情報については、図 3 を参照して後述する。

20

【0015】

構文情報検索部 106 は、造語候補検出部 103 から後述する校正制御部 107 を通じて造語候補文字列を受け取る。そして、構文情報検索部 106 は、造語候補文字列を構成する構成語が格要素となり、かつ構成語が名詞連続関係以外の依存関係を有する文（以下、依存関係文という）があるかどうかを、構文情報データベース 105 に格納されている構文情報から検索する。その後、構文情報検索部 106 は、依存関係文があれば造語候補文字列を造語であると判定する。格要素の具体的な判定については後述する。

【0016】

校正制御部 107 は、文書入力部 101 から文書の校正処理を行う場合には校正文書を受け取り、文書の登録処理を行う場合には登録文書を受け取る。校正処理を行う場合は、校正制御部 107 は、造語候補検出部 103 から造語候補文字列を受け取り、構文情報検索部 106 の要求に応じて造語候補文字列を構文情報検索部 106 に送る。また、校正制御部 107 は、構文情報検索部 106 において造語候補文字列が造語であると判定された場合は、構文情報検索部 106 から造語と依存関係文とを受け取る。さらに、校正制御部 107 は、後述する警告表示部 108 の要求に応じて造語と依存関係文とを警告表示部 108 に送る。

30

登録処理を行う場合は、校正制御部 107 は、文書解析部 102 から校正文書中の文とその解析情報とを受け取り、構文情報管理部 104 の要求に応じて文とその解析情報とを構文情報管理部 104 に送る。なお、登録処理を行う場合に入力される登録文書は、インターネット等に保持されている文書群から抽出してもよいし、ユーザが過去に作成した文書でも現在作成している文書の既に執筆済みの部分でもよい。

40

【0017】

警告表示部 108 は、構文情報検索部 106 から校正制御部 107 を通じて造語と依存関係文とを受け取り、警告メッセージとして表示する。警告メッセージの具体例については後述する。

【0018】

次に、本実施形態に係る文書校正支援装置の造語判定処理について図 2 のフローチャートを参照して詳細に説明する。なお、入力文書は日本語で書かれた文書とする。

【0019】

50

ステップS201では、文書解析部102が、校正文書中の各文に関して形態素解析を行い、解析情報を生成する。

【0020】

ステップS202では、造語候補検出部103が、解析情報を参照して造語候補文字列があるかどうかを検出する。ここでは、造語候補文字列は名詞またはサ変名詞の2以上の語が連続する文字列とする。このような造語候補文字列があればステップS203に進み、造語候補文字列がなければステップS205へ進む。

【0021】

ステップS203では、構文情報検索部106が、構文情報データベース105から造語候補文字列を構成する構成語がガ格、ヲ格、ノ格などの助詞の格要素となり、かつ構成語に関する依存関係文があるかどうかを検索する。依存関係文があれば、造語候補文字列は造語であるとしてステップS204へ進み、依存関係文がなければ、ステップS205へ進む。

10

【0022】

ステップS204では、警告表示部108が造語と依存関係文とを警告メッセージとして表示し、ユーザに造語が使用されていることを通知する。

【0023】

ステップS205では、校正制御部107が校正文書中の次の文を読み込み、次の文に対して上述した処理と同様の処理を行うためステップS201へ戻る。

【0024】

ここで、具体例を用いて本実施形態に係る文書校正支援装置の造語判定処理について説明する。

20

ここでは、以下の日本語文(1)を校正文書中に存在している文として造語判定処理をする場合を考える。

「崩壊熱は、燃料寿命末期において、炉停止してから1時間後で約0.4MWtである。」・・・(1)

この日本語文(1)において、「炉停止してから1時間後」という部分を翻訳する場合、この文書が示す技術内容を正確に翻訳するためには、「(誰か、もしくは何かが)炉を停止処理してから1時間後」であるのか、または「炉が停止してから1時間後」であるのかという単語間の意味的なつながりを知ることが重要である。しかし、原子力分野に詳しくないユーザが翻訳を行おうとした場合にユーザが技術内容を判断することは難しい。そこで、本実施形態に係る文書校正支援装置の造語判定処理を行う。

30

【0025】

まず、文書解析部102が、日本語文(1)について形態素解析をおこなった結果を(2)に示す。

崩壊熱(名詞)/は(助詞)/、/燃料寿命(名詞)/末期(名詞)/に(助詞)/お(動詞)/て(接続助詞)/、/炉(名詞)/停止(サ変名詞)/する(動詞)/て(接続助詞)/から(接続助詞)/1(数字)/時間(数量詞)/後(名詞)/で(助詞)/約(冠数詞)/0.4(数字)/MWt(英未)/であ(助動詞)/。(句点)・・・(2)

40

ここで、“/”は形態素の境界を表し、「見出し語語幹(品詞)」というフォーマットで記述している。

【0026】

次に、造語候補検出部103が、名詞またはサ変名詞の2以上の語が連続する複合語があるかどうかを検出する。(2)の例では、「炉(名詞)/停止(サ変名詞)」が該当するため、「炉(名詞)/停止(サ変名詞)」を造語候補文字列として検出する。

【0027】

次に、構文情報検索部106が、造語候補検出部103から校正制御部107を通じて受け取った造語候補文字列「炉停止」の構成語「炉」と「停止」とがガ格、ヲ格、ノ格などの助詞の格要素となっており、「炉」と「停止」とが文書中に互いに依存関係を有する

50

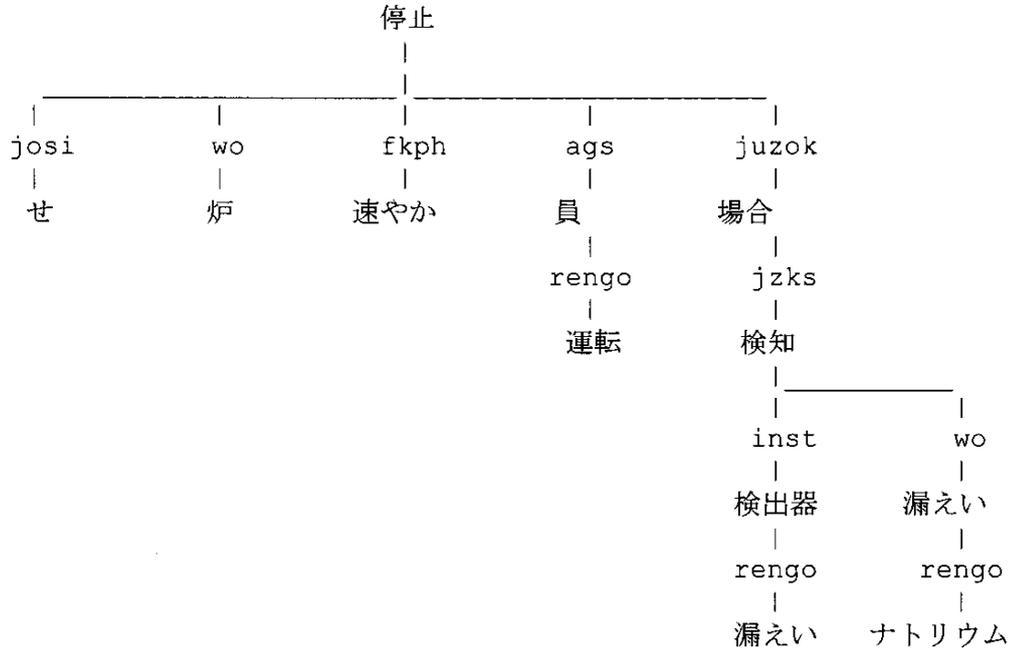
文（依存関係文）があるかどうかを構文情報データベース105から検索する。一例として構文情報データベース105に「炉」と「停止」とに関する依存関係文（3）が存在したとする。

「ナトリウム漏えいが漏えい検出器で検出された場合、運転員は炉を速やかに停止させる。」・・・（3）

さらに、構文情報データベース105に、表1に示すような依存関係文（3）の構文および係り受け情報が登録されていたとする。

【表1】

表1



- josi : 助動詞的句
- wo : 「ヲ」格
- fkph : 副詞句
- ags : 「させる」動詞表現の主体
- rengo : 名詞句中の名詞・形容動詞語幹・記号・数字などの連続を表す
- juzok : 従属副詞節
- jzks : 従属副詞節が従える節
- inst : 道具格

【0028】

表1は、構文および係り受け情報から格と見出し語とを用いて木構造で表現したものである。表1を参照すれば、「炉+ヲ(wo)+停止する」というヲ格による依存関係を検出することができる。従って、「炉停止」または「炉を停止する」という2種類の表現が存在することになる。すなわち、「炉停止」は「炉+ヲ(wo)+停止する」という表現を簡略化して名詞化した表現であり、「炉停止」は造語であると判定することができる。このように、構文解析および係り受け解析を行った解析情報を参照することにより、各構成語がどのような格要素によって互いに依存関係を有するかを解析することができる。

【0029】

なお、「炉停止」という造語は、「炉(を)停止する」というヲ格による依存関係だけでなく、「炉(が)停止する」というガ格による依存関係を有する可能性もある。そこで、「炉停止」が造語であると判定された場合、校正制御部107は造語と依存関係文と含んだ警告メッセージをユーザに通知する。警告メッセージの一例を表2に示す。

## 【表 2】

表 2

造語が検出されました：「炉停止」  
過去の用法：「ナトリウム漏えいが漏えい検出器で検知された場合、運転員は炉を速やかに停止させる。」

## 【 0 0 3 0 】

ここで表 2 中の下線は造語を構成する構成語が、依存関係文中でどのように用いられているかを示す。表 2 に示す警告メッセージであれば、ユーザは、造語として検出された複  
10  
合語は過去にどのような依存関係で用いられていたかを認識することができ、造語の内容を正確に把握した上で翻訳することができる。なお、構成語を下線で表示するだけでなく、太字、ハイライト、または点滅させることによって表示してもよい。また、警告メッセージの表示方法は、ユーザが認識できる方法であれば何でもよく、ユーザが作業している画面中に別ウィンドウで表示してもよいし、文書中にポップアップで表示してもよい。さらに、警告メッセージを文書でユーザに提示するだけでなく、例えば、音声によりユーザに通知してもよい。

## 【 0 0 3 1 】

次に、構文情報データベース 1 0 5 に格納される構文情報の一例について図 3 を参照して  
20  
詳細に説明する。

構文情報データベース 1 0 5 には、ID 3 0 1、形態素情報 3 0 2、構文および係り受け情報 3 0 3、原文 3 0 4、および校正情報 3 0 5 が構文情報としてそれぞれ対応付けて格納される。図 3 では、「ナトリウム漏えいが漏えい検出器で検出された場合、運転員は炉を速やかに停止させる」という原文 3 0 4 ごとにユニークな ID 3 0 1 が示される。さらに、原文 3 0 4 を形態素解析した形態素情報 3 0 2、原文 3 0 4 を構文解析および係り受け解析した構文および係り受け情報 3 0 3、および原文 3 0 4 が校正済の文であるかどうかを示す校正情報 3 0 5 が示される。

## 【 0 0 3 2 】

校正情報 3 0 5 は、図 3 では、原文が校正済である状態を「済」、原文がまだ校正されてない状態を「未」とする。なお、これに限らず、フラグが立っている場合を校正済とする  
30  
手法により校正済であるかどうかを表現してもよく、校正済の文であるかどうかを判別できる手法であれば何でもよい。校正情報 3 0 5 が「済」である場合、造語候補文字列がこの校正済の文中に存在していれば、文中に存在する造語候補文字列は過去に 1 度以上造語判定処理が行われたと考えられるため、造語として表示しないようにする。こうすることで、過去の校正済の文書資源を有効に利用することができる。

なお、構文情報は、文書が作成された日付を示す情報や執筆者の情報などを付加し、警告表示部 1 0 8 においてさらに詳細な情報を表示できるようにしてもよい。

## 【 0 0 3 3 】

次に、構文情報検索部 1 0 6 が単語間の依存関係を検索する一例について図 4 を参照して  
40  
説明する。

構文情報検索部 1 0 6 が造語候補文字列の構成語間の依存関係を検索する場合に、本実施形態では、構文情報データベース 1 0 5 に格納されたインデックステーブルを用いる。インデックステーブルは、形態素情報から作成したインデックス 4 0 1 と図 3 に示す ID 3 0 1 とをそれぞれ対応させたテーブルである。インデックステーブルの作成については、例えば、登録文書中の文に対して構文情報が作成された後、構文情報管理部 1 0 4 が構文情報から形態素情報 3 0 2 と ID 3 0 1 とを抽出して作成すればよい。例えば、「炉」と「停止」との単語間の依存関係を検索する場合は、「炉」と「停止」とのそれぞれについてインデックス 4 0 1 を検索し、共通の ID 3 0 1 を有していれば単語間の依存関係を得ることができる。この場合では、ID 3 0 1 「0 0 1」が該当し、図 3 に示すような構文情報中の構文および係り受け情報 3 0 3 を参照することで単語間の格要素による依存関  
50

係を得ることができる。

【0034】

(本実施形態の変形例)

ここでは、本実施形態の変形例として、日本語による文書ではなく、中国語による文書の場合の本実施形態に係る文書校正支援装置について説明する。

【0035】

一例として、入力された校正文書中に中国語文(4)が存在していると仮定する。

【表3】

「AB Enterprise OS 启动模式変更」・・・(4)

10

【0036】

この文中の中国語には(5)に示すように、2通りの解釈が存在する。

【表4】

『模式変更(モード変更)』の『启动(起動)』

または

(5)

『启动模式(起動モード)』の『変更(変更)』

【0037】

従って、上述の日本語技術文書の例のときと同様に、この技術分野の専門家でない翻訳者がこの文を英語などの中国語以外の言語に翻訳する際に困難を伴うことが予想される。

20

このような複合語を造語として検出したい場合、例えば従来の手法のようにコーパス出現頻度を調べ、その頻度が閾値以下であれば造語とする手法が考えられる。しかし、中国語の場合は単に複合語としてではなく、「動詞+名詞」または「名詞+動詞」として文中に存在する場合があるので、この単語列の頻度だけでは、複合語が造語であるかどうかの判定は難しい。

【0038】

または、以下の中国語(6)から中国語(9)までに示す中国語の部分単語列のコーパス出現頻度を調べ、その頻度が高いものを部分複合語とし、中国語文(4)が中国語(6)から中国語(9)まで以外の単語との連結によって生成された造語であると判定する手法も考えられる。

30

【表5】

「模式変更」・・・(6)

「启动模式」・・・(7)

「启动」・・・・・・(8)

「変更」・・・・・・(9)

【0039】

しかし、この例文の技術分野であるコンピュータ分野では、これらの単語はコーパス出現頻度が高く、この手法では複合語が造語であるかどうかの判定は難しい。

40

【0040】

ここで、中国語の文書の場合における本実施形態に係る文書校正支援装置の造語判定処理について、図5のフローチャートを参照して詳細に説明する。ここでは、中国語文(4)を入力文として造語判定処理を行う。

ステップS501では、文書解析部102が入力文の形態素解析を行う。

【0041】

ステップS502では、造語候補検出部103が、造語候補文字列があるかどうかを検出する。本変形例の場合の造語候補文字列は、名詞または名詞および動詞の両解釈がある単語の2以上の語が連続する複合語とする。このような造語候補文字列があればステップ

50

S 5 0 3へ進み、造語候補文字列がなければステップS 5 0 5へ進む。具体例として、造語候補検出部 1 0 3 が上記中国語文( 4 )から造語候補文字列を検出する場合を考える。

中国語( 8 )は、「起動( 名詞 )」と「起動する( 動詞 )」との両方の解釈がある。次の「モード」は名詞であり、中国語( 9 )は、「変更( 名詞 )」と「変更する( 動詞 )」との両方の解釈がある。

【 0 0 4 2 】

従って、このように 2 以上の語が連続する複合語を造語候補文字列( 1 0 )として検出する。

【表 6】

10

「启动模式变更」・・・( 1 0 )

【 0 0 4 3 】

ステップS 5 0 3では、構文情報検索部 1 0 6 が構文情報データベース 1 0 5 を検索して、造語候補文字列の構成語が動詞の格要素となり、かつ構成語に関する依存関係文があるかどうかを検索する。依存関係文があれば、造語候補文字列は造語であるとしてステップS 5 0 4に進み、依存関係文がなければ、ステップS 5 0 5へ進む。

【 0 0 4 4 】

構文情報データベース 1 0 5 に次のような中国語文( 1 1 )とその解析結果である構文情報が格納されている場合を考える。

20

【表 7】

请按照说明书把启动模式变更为自动。・・・( 1 1 )

(訳：マニュアルに従って起動モードを自動に変更してください。)

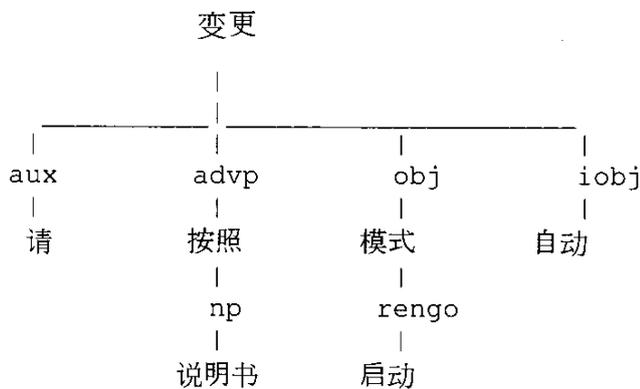
【 0 0 4 5 】

中国語文( 1 1 )の下線部は校正対象の中国語文( 4 )に存在する造語候補文字列と同一である。また、中国語文( 1 1 )の構文および係り受け解析情報を表 3 に示す。

【表 8】

30

表 3



40

- aux : 助動詞
- advp : 動詞句・形容詞句を修飾する前置句
- np : 名詞句
- obj : 動詞の必須格である目的語
- iobj : 動詞の必須格である間接目的語
- rengo : 名詞句中の名詞、形容動詞語幹、記号、および数字などの連続を表す

【 0 0 4 6 】

表 3 の構文および係り受け解析情報を参照すると、主動詞は中国語( 9 )であり、中国

50

語(7)は目的語(obj)の格要素であることが分かる。よって、中国語(9)は動詞の格要素であり、中国語(7)と中国語(9)との間には「目的語 動詞」の依存関係が存在するため造語候補文字列(10)は造語であると判定する。

【0047】

ステップS504では、警告表示部108が造語と依存関係文とを表示してユーザに造語が使用されてことを通知する。警告表示部108が表示する警告メッセージの一例を表4に示す。

【表9】

表 4

被查出新词了：「启动模式变更」 过去的用法：「请按照说明书把启动模式变更为自动。」
--

10

【0048】

警告メッセージの表示方法は、日本語文書における造語判定処理と同様に、ユーザが認識できる方法であれば何でもよい。

【0049】

以上に示した実施形態によれば、造語候補文字列を構成する構成語に対して、過去の登録文書中の文に存在する構成語間の依存関係により造語かどうかを判定するので、原言語のコーパス頻度を利用した造語の検出および指摘よりも精度のよい造語の検出および指摘が可能である。さらに、過去の登録文書中の文において構成単語が同一文中に存在していたとしても係り受け関係がない場合は、対象の複合語は造語とは認識しないなど、文字列検索またはワイルドカード検索よりも柔軟で精度のよい指摘ができる。よって、文書校正の精度を高め作業効率を向上させることができる。

20

【0050】

また、上述の実施形態の中で示した処理手順に示された指示は、ソフトウェアであるプログラムに基づいて実行されることが可能である。汎用の計算機システムが、このプログラムを予め記憶しておき、このプログラムを読み込むことにより、上述した文書校正支援装置による効果と同様な効果を得ることも可能である。上述の実施形態で記述された指示は、コンピュータに実行させることのできるプログラムとして、磁気ディスク(フレキシブルディスク、ハードディスクなど)、光ディスク(CD-ROM、CD-R、CD-RW、DVD-ROM、DVD±R、DVD±RWなど)、半導体メモリ、又はこれに類する記録媒体に記録される。コンピュータまたは組み込みシステムが読み取り可能な記録媒体であれば、その記憶形式は何れの形態であってもよい。コンピュータは、この記録媒体からプログラムを読み込み、このプログラムに基づいてプログラムに記述されている指示をCPUで実行させれば、上述した実施形態の文書校正支援装置と同様な動作を実現することができる。もちろん、コンピュータがプログラムを取得する場合又は読み込む場合はネットワークを通じて取得又は読み込んでよい。

30

また、記録媒体からコンピュータや組み込みシステムにインストールされたプログラムの指示に基づきコンピュータ上で稼働しているOS(オペレーティングシステム)や、データベース管理ソフト、ネットワーク等のMW(ミドルウェア)等が本実施形態を実現するための各処理の一部を実行してもよい。

40

さらに、本願発明における記録媒体は、コンピュータあるいは組み込みシステムと独立した媒体に限らず、LANやインターネット等により伝達されたプログラムをダウンロードして記憶または一時記憶した記録媒体も含まれる。

また、記録媒体は1つに限られず、複数の媒体から本実施形態における処理が実行される場合も、本発明における記録媒体に含まれ、媒体の構成は何れの構成であってもよい。

【0051】

なお、本願発明におけるコンピュータまたは組み込みシステムは、記録媒体に記憶されたプログラムに基づき、本実施形態における各処理を実行するためのものであって、パソ

50

コン、マイコン等の1つからなる装置、複数の装置がネットワーク接続されたシステム等の何れの構成であってもよい。

また、本願発明の実施形態におけるコンピュータとは、パソコンに限らず、情報処理機器に含まれる演算処理装置、マイコン等も含み、プログラムによって本発明の実施形態における機能を実現することが可能な機器、装置を総称している。

【0052】

なお、本発明は上記実施形態そのままに限定されるものではなく、実施段階ではその要旨を逸脱しない範囲で構成要素を変形して具体化できる。また、上記実施形態に開示されている複数の構成要素の適宜な組み合わせにより、種々の発明を形成できる。例えば、実施形態に示される全構成要素から幾つかの構成要素を削除してもよい。さらに、異なる実施形態にわたる構成要素を適宜組み合わせてもよい。

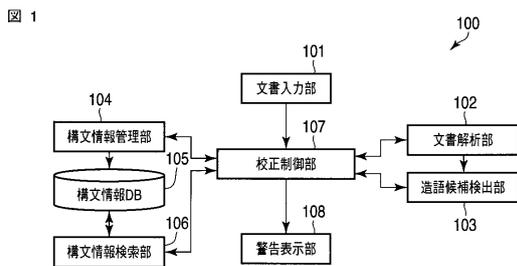
【符号の説明】

【0053】

100・・・文書校正支援装置、101・・・文書入力部、102・・・文書解析部、103・・・造語候補検出部、104・・・構文情報管理部、105・・・構文情報データベース(DB)、106・・・構文情報検索部、107・・・校正制御部、108・・・警告表示部、301・・・ID、302・・・形態素情報、303・・・構文および係り受け情報、304・・・原文、305・・・校正情報、401・・・インデックス。

10

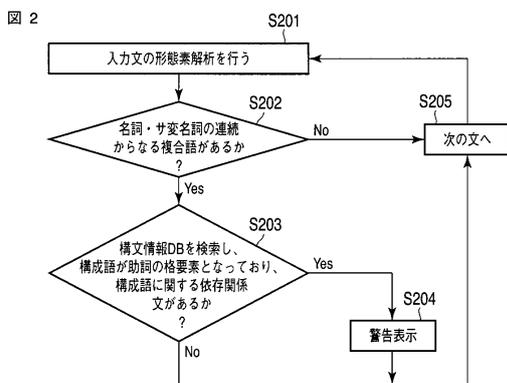
【図1】



【図3】

ID	形態素情報	構文および係り受け情報
	001	ナトリウム(名詞)/漏えい(サ変名詞)/が(助詞)/漏えい(サ変名詞)/検出器(名詞)/で(助詞)/検出(サ変名詞)/き(助詞)/れ(助動詞)/た(助動詞)/場合(接尾的名詞)/、/運転(名詞)/員(接尾)/は(助詞)/炉(名詞)/を(助詞)/速やか(形容動詞)/停止(サ変名詞)/き(助詞)/せ(助動詞)。(句点)
	原文	校正情報
	ナトリウム漏えいが漏えい検出器で検出された場合、運転員は炉を速やかに停止させる	未

【図2】

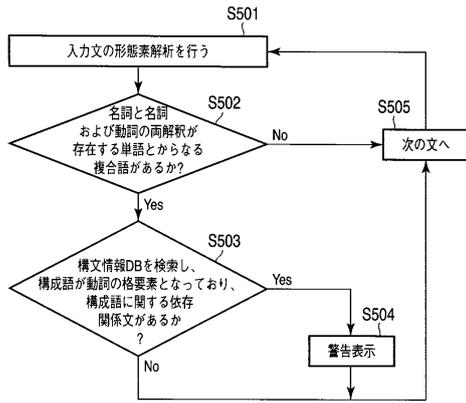


【図4】

インデックス	ID
ナトリウム	001/004/010...
漏えい	001/007/...
...	...
炉	001/023
停止	001/002/058

【 図 5 】

図 5



## フロントページの続き

- (74)代理人 100084618  
弁理士 村松 貞男
- (74)代理人 100103034  
弁理士 野河 信久
- (74)代理人 100119976  
弁理士 幸長 保次郎
- (74)代理人 100153051  
弁理士 河野 直樹
- (74)代理人 100140176  
弁理士 砂川 克
- (74)代理人 100101812  
弁理士 勝村 紘
- (74)代理人 100124394  
弁理士 佐藤 立志
- (74)代理人 100112807  
弁理士 岡田 貴志
- (74)代理人 100111073  
弁理士 堀内 美保子
- (74)代理人 100134290  
弁理士 竹内 将訓
- (74)代理人 100127144  
弁理士 市原 卓三
- (74)代理人 100141933  
弁理士 山下 元
- (72)発明者 鈴木 博和  
東京都港区芝浦一丁目1番1号 株式会社東芝内

審査官 成瀬 博之

- (56)参考文献 特開2007-241764(JP,A)  
乾健太郎 他1名, 言い換え技術に関する研究動向, 自然言語処理, 日本, 言語処理学会, 2004年10月10日, 第11巻第5号, 151-198頁, 特に「5.3.2 内部構造の類似性を測る」の項目参照  
久光徹, 文書走査を用いた複合名詞解析について, 情報処理学会研究報告, 日本, 社団法人情報処理学会, 1996年3月15日, Vol196, No.27, 7-14頁

- (58)調査した分野(Int.Cl., DB名)  
G06F 17/20 - 17/28