(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2011-123794
(P2011-123794A)
(43)公開日　平成23年6月23日(2011.6.23)

(51)Int.Cl.		F I			テーマコード（参考）
G06F 17/27	(2006.01)	G06F	17/27	Z	5B075
G06F 17/30	(2006.01)	G06F	17/30	220Z	5B091
		G06F	17/30	170A	

審査請求 有　請求項の数 7　ＯＬ　（全 24 頁）

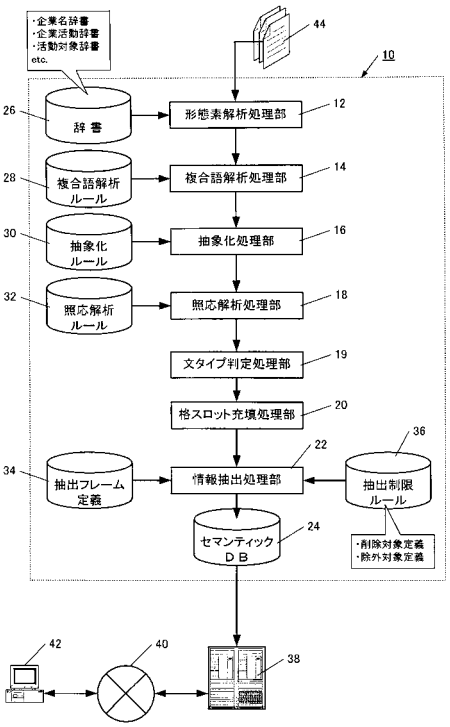
(21)出願番号　特願2009-282686 (P2009-282686)
(22)出願日　平成21年12月14日 (2009.12.14)
(11)特許番号　特許第4625535号 (P4625535)
(45)特許公報発行日　平成23年2月2日 (2011.2.2)

(71)出願人　000155469
株式会社野村総合研究所
東京都千代田区丸の内一丁目６番５号
(74)代理人　100096002
弁理士　奥田　弘之
(74)代理人　100091650
弁理士　奥田　規之
(72)発明者　竹原　一彰
東京都千代田区丸の内一丁目６番５号　株式会社野村総合研究所内
(72)発明者　大島　修
東京都千代田区丸の内一丁目６番５号　株式会社野村総合研究所内

最終頁に続く

(54)【発明の名称】情報抽出システム及び情報抽出プログラム

(57)【要約】
【課題】構文解析技術を用いることなく、自然文から構造化された情報を抽出する技術の提供。
【解決手段】企業名、企業活動、活動対象物を示す具体的な表現文字列毎にその種類を示す抽象化文字列を登録した辞書記憶部26と、文を形態素単位に分解し、各形態素に対応の抽象化タグを関連付ける形態素解析処理部12と、企業活動の抽象化タグが付与された形態素を文の述語と認定すると共に、主語に付属する助詞毎及び目的語に付属する助詞毎に対応語の格納欄が設けられた格スロットに、文の述語単位で対応語を充填し、述語を関連付ける格スロット充填処理部20と、抽出すべき主語の抽象化タグ及び助詞を特定する条件と、抽出すべき述語の抽象化タグを特定する条件と、抽出すべき目的語の抽象化タグ及び助詞を特定する条件が規定された抽出フレーム定義を、対応語充填済みの格スロットに適用し、文の主語、述語、目的語に該当する情報要素を抽出する情報抽出処理部22を備えた情報抽出システム10。
【選択図】図１



【特許請求の範囲】
【請求項１】
活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、
上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、
上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、
テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段と、
上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には、当該形態素に対してその種類を示す抽象化タグを関連付ける手段と、
少なくとも主語に付属する助詞毎及び目的語に付属する助詞毎に対応語の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化タグが付与されている形態素を述語として当該格スロットに関連付ける格スロット充填手段と、
抽出すべき主語の抽象化タグ及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化タグを特定する条件と、抽出すべき目的語の抽象化タグ及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段と、
対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する情報要素を抽出する情報抽出手段とを備えた情報抽出システムであって、
上記格スロット充填手段が、以下の処理を実行することを特徴とする情報抽出システム。
(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。
(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。
(3)文の主語を表す助詞の対応語格納欄が後続の述語に係る有意思主体を表す語によって上書充填された場合には、先行する述語に関して充填された対応語を削除する。
【請求項２】
上記の文の中でタイトルに該当する文に対して、タイトル文であることを示す識別情報を予め付与する手段を備え、
この識別情報が付与されたタイトル文に対して、上記格スロット充填手段が以下の処理を実行することを特徴とする請求項１に記載の情報抽出システム。
(1)タイトル文中の主語となるべき種類の抽象化タグが付与された語については、助詞の有無を問わず主語に付属する助詞の対応語格納欄に充填する。
(2)タイトル文中の目的語となるべき種類の抽象化タグが付与された語については、助詞の有無を問わず目的語に付属する助詞の対応語格納欄に充填する。
【請求項３】
文中における述語の前に目的語に付属する助詞が存在しない場合に、倒置表現文であることを示す識別情報を予め付与する手段を備え、
この識別情報が付与されたタイトル文に対して、上記格スロット充填手段が当該述語に後続する名詞を文の目的語と認定し、格スロットの目的語に付属する助詞の対応語格納欄に当該名詞を充填することを特徴とする請求項１または２に記載の情報抽出システム。
【請求項４】
複合語となるべき複数の品詞の連結パターン毎に、当該複合語の品詞を決定するための基準が規定された複合語解析ルールを格納しておく複合語解析ルール記憶手段と、
この複合語解析ルールを参照し、文中に複合語解析ルールに規定された品詞の連結パターンに該当する形態素の組合せが存在している場合には、これらの形態素を複合語と認定する複合語解析手段とを備え、
上記の格スロット充填手段は、複合語と認定された形態素の組合せについては、複合語

単位で格スロットへの充填処理を実行することを特徴とする請求項１～３の何れかに記載の情報抽出システム。
【請求項５】
形態素の種類を推定するための抽象化ルールを格納しておく抽象化ルール記憶手段と、
上記の抽象化ルールを文に対して適用し、当該抽象化ルールにマッチする形態素に対してその種類を示す抽象化タグを関連付ける手段と、
を備えたことを特徴とする請求項１～４の何れかに記載の情報抽出システム。
【請求項６】
照応詞毎に、その先行詞を決定するための基準を定めた照応解析ルールを格納しておく照応解析ルール記憶手段と、
この照応解析ルールを参照し、文中に存する照応詞に対して、対応の先行詞を決定すると共に、この先行詞によって照応詞を置き換える照応解析手段とを備えたことを特徴とする請求項１～５の何れかに記載の情報抽出システム。
【請求項７】
コンピュータを、
活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、
上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、
上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、
テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段、
上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には、当該形態素に対してその種類を示す抽象化タグを関連付ける手段、
少なくとも主語に付属する助詞毎及び目的語に付属する助詞毎に対応語の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化タグが付与されている形態素を述語として当該格スロットに関連付ける格スロット充填手段、
抽出すべき主語の抽象化タグ及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化タグを特定する条件と、抽出すべき目的語の抽象化タグ及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段、
対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する情報要素を抽出する情報抽出手段として機能させる情報抽出プログラムであって、
上記格スロット充填手段が、以下の処理を実行することを特徴とする情報抽出プログラム。
(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。
(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。
(3)文の主語を表す助詞の対応語格納欄が後続の述語に係る有意思主体を表す語によって上書充填された場合には、先行する述語に関して充填された対応語を削除する。
【発明の詳細な説明】
【技術分野】
【０００１】
この発明は情報抽出システム及び情報抽出プログラムに係り、特に、構造化されていないテキストデータから各企業の活動内容や活動対象を定義した多数の企業情報等を自動抽出する技術に関する。
【背景技術】
【０００２】
インターネット上のニュースサイトにおいて公開されているWebページなど、自然言語

で記述された構造化されていないテキストデータから必要な情報を抽出するための技術として、構文解析技術を用いるものが種々提案されている。

例えば、特許文献１に記載の情報抽出装置の場合、自然言語で記述された文書中の文字列と所定の文字パターンとを逐次照合し、一致が認められた文字列部分に対し固有名詞の種類を示すタグ情報を付与する文字パターン処理部と、上記タグ情報はそのままに、タグ情報を除く他の文字列部分を逐次単語情報に分割する形態素解析処理部と、形態素解析の結果得られた単語情報を文節単位にまとめ上げ、当該まとめ上げ後の単語情報を、文法上の構文規則と共に、ある種の情報の表現に特徴的に現れる構文パターンを用いて構文解析する構文解析部と、上記構文パターンに基づく解析により得られる係り受け関係及び当該係り受け関係に含まれるタグ情報から特定される情報を、必要な情報として抽出する情報抽出部を備えている。

【特許文献１】特開平１１－２７２６９５号

【発明の開示】

【発明が解決しようとする課題】

【０００３】

この従来の情報抽出装置を用いることにより、例えば「５日午前零時３５分ごろ、大阪市中央町、消毒業、鈴木勇さん（５０）方から出火、木造平屋建て約１２５平方メートルが全焼した。」という文章から、「＜人名＞鈴木勇さん｜＜地名＞大阪市中央町｜＜業種名＞消毒業」の構造化された情報が抽出可能となる。

【０００４】

しかしながら、このように構文解析処理を前提とした情報抽出方式の場合、必要な情報を正確に抽出するためには、抽出対象となる形態素を指定するための構文パターンを多数準備しておく必要があり、そのために多大なコストを要していた。

【０００５】

すなわち、自然文の場合には単純な文章ばかりでなく、複雑な構文構造を備えたものが多いため、その中から必要な情報を抽出するためには、多様な構文パターンを用意する必要がある。

【０００６】

例えば、「東洋自動車子会社の変速機メーカーパイロンは、燃費効率の高い変速機である無段変速機を中国の広州市で生産すると発表した」という文章の場合、複数の企業名（東洋自動車、パイロン）、複数の対象物（変速機、無段変速機）、及び複数の活動内容（生産、発表）が形式上含まれているため、ここから真の活動主体（主語）、活動対象（目的語）、活動内容（述語）を抽出するとなると、図１５に示すように、まず文全体を構文解析して文節間の係り受け構造を明らかにし、各種辞書を参照して各文節に種類を表すタグ（＜企業名＞等）を付与した後、図１６に示すように、抽出すべき形態素の種類及び文中の位置関係を定義した構文パターンＡを用意する必要がある。

この構文パターンＡを図１５の構文解析結果に適用することにより、図１７(a)に示す文節が文中より抽出され、これに必要な整形処理を施すことにより、図１７(b)に示すように、主語、述語、目的語の組合せからなる構造化された企業情報が得られる。

【０００７】

また、文が「東洋自動車子会社の変速機メーカーパイロンは、燃費効率の高い変速機である無段変速機を中国の広州市で生産する」と表現される場合を想定し、図１８に示すように、別の構文パターンＢを用意しておく必要がある。

【０００８】

さらに、自然文の場合には修辞上の目的で倒置表現や省略表現が多く用いられるが、倒置表現を含む文章からも必要な情報を抽出するためには、倒置表現を前提とした構文パターンを事前に多数用意しておく必要があった。省略表現にいたっては、そもそも省略されている要素を構文パターンとして定義することができないため、省略表現を含む文章からの情報抽出自体が不可能であった。

【０００９】

この発明は、従来のこのような問題を解決するために案出されたものであり、文節間の係り受け構造に基づく構文解析処理を行うことなく、したがって情報を抽出するための構文パターンを用いることなく、自然文で記述された文章から構造化された情報を抽出可能な技術の提供を目的としている。

【課題を解決するための手段】

【００１０】

上記の目的を達成するため、請求項１に記載した情報抽出システムは、活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段と、上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には、当該形態素に対してその種類を示す抽象化タグを関連付ける手段と、少なくとも主語に付属する助詞（「は」、「が」）毎及び目的語に付属する助詞（「を」、「に」等）毎に対応語（自立語）の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化タグが付与されている形態素を述語として当該格スロットに関連付ける格スロット充填手段と、抽出すべき主語の抽象化タグ及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化タグを特定する条件と、抽出すべき目的語の抽象化タグ及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段と、対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する情報要素を抽出する情報抽出手段とを備えた情報抽出システムであって、上記格スロット充填手段が、以下の処理を実行することを特徴としている。

(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。

(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。

(3)文の主語を表す助詞の対応語格納欄が後続の述語に係る有意思主体を表す語によって上書充填された場合には、先行する述語に関して充填された対応語を削除する。

【００１１】

請求項２に記載した情報抽出システムは、請求項１のシステムであって、さらに、上記の文の中でタイトルに該当する文に対して、タイトル文であることを示す識別情報を予め付与する手段を備え、この識別情報が付与されたタイトル文に対して、上記格スロット充填手段が以下の処理を実行することを特徴としている。

(1)タイトル文中の主語となるべき種類の抽象化タグが付与された語については、助詞の有無を問わず主語に付属する助詞（「は」、「が」）の対応語格納欄に充填する。

(2)タイトル文中の目的語となるべき種類の抽象化タグが付与された語については、助詞の有無を問わず目的語に付属する助詞（「を」、「に」等）の対応語格納欄に充填する。

【００１２】

請求項３に記載した情報抽出システムは、請求項１または２のシステムであって、さらに、文中における述語の前に目的語に付属する助詞（例えば「を」）が存在しない場合に、倒置表現文であることを示す識別情報を予め付与する手段を備え、この識別情報が付与されたタイトル文に対して、上記格スロット充填手段が当該述語に後続する名詞を当該述語の目的語と認定し、格スロットの目的語に付属する助詞の対応語格納欄に当該名詞を充填することを特徴としている。

【００１３】

請求項４に記載した情報抽出システムは、請求項１～３のシステムであって、さらに、複合語となるべき複数の品詞の連結パターン毎に、当該複合語の品詞を決定するための基準が規定された複合語解析ルールを格納しておく複合語解析ルール記憶手段と、この複合

語解析ルールを参照し、文中に複合語解析ルールに規定された品詞の連結パターンに該当する形態素の組合せが存在している場合には、これらの形態素を複合語と認定する複合語解析手段とを備え、上記の格スロット充填手段は、複合語と認定された形態素の組合せについては、複合語単位で格スロットへの充填処理を実行することを特徴としている。

【００１４】

請求項５に記載した情報抽出システムは、請求項１～４のシステムであって、さらに、形態素の種類を推定するための抽象化ルールを格納しておく抽象化ルール記憶手段と、上記の抽象化ルールを文に対して適用し、当該抽象化ルールにマッチする形態素に対してその種類を示す抽象化タグを関連付ける手段とを備えたことを特徴としている。

【００１５】

請求項６に記載した情報抽出システムは、請求項１～５のシステムであって、さらに、照応詞（代名詞等）毎にその先行詞を決定するための基準を定めた照応解析ルールを格納しておく照応解析ルール記憶手段と、この照応解析ルールを参照し、文中に存する照応詞に対して、対応の先行詞を決定すると共に、この先行詞によって照応詞を置き換える照応解析手段とを備えたことを特徴としている。

【００１６】

請求項７に記載した情報抽出プログラムは、コンピュータを、活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段、上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には当該形態素に対してその種類を示す抽象化タグを関連付ける手段、少なくとも主語に付属する助詞毎及び目的語に付属する助詞毎に対応語の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化タグが付与されている形態素を述語として当該格スロットに関連付ける格スロット充填手段、抽出すべき主語の抽象化タグ及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化タグを特定する条件と、抽出すべき目的語の抽象化タグ及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段、対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する情報要素を抽出する情報抽出手段として機能させる情報抽出プログラムであって、上記格スロット充填手段が、以下の処理を実行することを特徴としている。

(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。

(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。

(3)文の主語を表す助詞の対応語格納欄が後続の述語に係る有意思主体を表す語によって上書充填された場合には、先行する述語に関して充填された対応語を削除する。

【発明の効果】

【００１７】

請求項１に記載した情報抽出システム及び請求項７に記載した情報抽出プログラムにあっては、予め必要な助詞の種類が設定された定型的な格スロットと、抽出すべき情報の種類を規定する抽出フレーム定義を用意しておき、文中に対応の助詞が存在する場合にはその直前の自立語を当該助詞の対応語格納欄に充填すると共に、この充填済みの格スロットに抽出フレーム定義を適用することにより、語順にかかわらず自然文から「主語－述語－目的語」のように構造化された情報を確実に抽出することができる。また、語順に拘束されないため、抽出フレーム定義のバリエーションを抑制することができる。

しかも、原則として前の述語に係る格スロットがつぎの述語に継承される仕組みを備えているため、後続の述語に関して主語や目的語の省略が存在したとしても、前の述語の主語や目的語で容易に補うことができる。

【００１８】
請求項２に記載した情報抽出システムによれば、助詞が省略されている場合が多いタイトル文に対しても、上記の格スロットを適用し、必要な語を抽出することが可能となる。
【００１９】
請求項３に記載した情報抽出システムによれば、文において目的語が述語の後ろに配置される倒置表現が用いられている場合であっても、的確に目的語を抽出することが可能となる。
【００２０】
請求項４に記載した情報抽出システムによれば、複数の形態素の組合せよりなる複合語を文中において的確に認定可能となり、この結果、情報要素を最適な粒度で抽出可能となる。
【００２１】
請求項５に記載した情報抽出システムによれば、辞書に収録されていない形態素についてもルールベースで抽象化タグを付与することが可能となり、その分、多くの情報要素をテキストから抽出可能となる。
【００２２】
請求項６に記載した情報抽出システムによれば、文中の照応詞を対応の先行詞で置き換えることが可能となり、その分、多くの情報要素をテキストから抽出可能となる。
【発明を実施するための最良の形態】
【００２３】
図１は、この発明に係る情報抽出システム10の全体構成を示すブロック図であり、形態素解析処理部12と、複合語解析処理部14と、抽象化処理部16と、照応解析処理部18と、文タイプ判定処理部19と、格スロット充填処理部20と、情報抽出処理部22と、セマンティックＤＢ24と、辞書記憶部26と、複合語解析ルール記憶部28と、抽象化ルール記憶部30と、照応解析ルール記憶部32と、抽出フレーム定義記憶部34と、抽出制限ルール記憶部36を備えている。
セマンティックＤＢ24には、検索サーバ38が接続されており、通信ネットワーク40を介して接続されたクライアント端末42に対し検索サービスを提供する。
【００２４】
上記の形態素解析処理部12、複合語解析処理部14、抽象化処理部16、照応解析処理部18、文タイプ判定処理部19、格スロット充填処理部20、情報抽出処理部22は、コンピュータのCPUが、OS及びアプリケーションプログラムに従って必要な処理を実行することによって実現される。
また、上記のセマンティックＤＢ24、辞書記憶部26、複合語解析ルール記憶部28、抽象化ルール記憶部30、照応解析ルール記憶部32、抽出フレーム定義記憶部34、抽出制限ルール記憶部36は、同コンピュータのハードディスク内に設けられている。
【００２５】
辞書記憶部26内には、企業名辞書、企業活動辞書、活動対象物辞書、人物名辞書、国名辞書、地域名辞書、都道府県名辞書、市町村名辞書、動植物名辞書、時間表現辞書、同義語辞書等が格納されている。
【００２６】
図２は、企業活動辞書の登録内容を例示するものであり、企業活動の一種である上位概念的な「生産活動」の抽象化文字列に対して、「生産」、「製造」、「加工」、「組立」等の述語となるべき具体的な表現文字列が予め対応付けられている。同じく、企業活動の一種である上位概念的な「販売活動」の抽象化文字列に対しては、「販売」、「発売」、「売り出す」等の述語となるべき具体的な表現文字列が予め対応付けられている。さらに、企業活動の一種である上位概念的な「開発活動」の抽象化文字列に対しては、「開発」、「研究」、「研究開発」等の述語となるべき具体的な表現文字列が予め対応付けられている。なお、「生産活動」や「販売活動」、「開発活動」の代わりに、より上位概念的な「企業活動」の抽象化文字列を用いて一まとめにしてもよい。

【００２７】
図３は、活動対象物辞書の登録内容を例示するものであり、上位概念的な「生産対象物」、「販売対象物」、「開発対象物」等の抽象化文字列に対して、「液晶」、「液晶テレビ」、「液晶パネル」、「液晶モニター」等の目的語となるべき具体的な表現文字列が予め対応付けられている。なお、「生産対象物」「販売対象物」、「開発対象物」等の代わりに、より上位概念的な「活動対象物」の抽象化文字列を用いてもよい。
【００２８】
図示は省略したが、企業名辞書には、主語となるべき具体的な企業名（正式名称及び略称）が、「企業名」の抽象化文字列に関連付けられて多数登録されている。
【００２９】
つぎに、このシステム10による処理内容を説明する。
まず、形態素解析処理部12により、外部から入力されたWebファイル等のテキストデータ44に対する形態素解析が実行される。ここで「形態素解析」とは、自然言語で記述された文を、意味を有する最小の言語単位である形態素に分解し、それぞれの品詞を同定する処理をいう。
【００３０】
例えば、「東洋自動車子会社の変速機メーカー、パイロンは２８日、燃費効率の高い変速機である無段変速機を中国の広州市で生産すると発表した。」という文章が与えられた場合、形態素解析処理部12はこれを「東洋（名詞-一般）／自動車（名詞-一般）／子会社（名詞-一般）／の（助詞-連体化）／変速（名詞-サ変接続）／機（名詞-接尾）／メーカー（名詞-一般）／、（記号-読点）／パイロン（名詞-一般）／…／広州（名詞-固有名詞）／市（名詞-接尾）／で（助詞-格助詞）／生産（名詞-サ変接続）／する（動詞-自立）／と（助詞-格助詞）／発表（名詞-サ変接続）／し（動詞-自立）／た（助動詞）／。（記号-句点）」のように分解し、それぞれの品詞を特定する。
この形態素解析自体は公知技術であり、例えば以下のようなフリーソフトを形態素解析エンジンとして用いることができる。
(1) MeCab（http://mecab.sourceforge.net/）
(2) ChaSen（http://chasen.naist.jp/hiki/ChaSen/）
【００３１】
つぎに形態素解析処理部12は、辞書記憶部26内に格納された企業名辞書、企業活動辞書、活動対象物辞書を参照し、特定形態素の品詞に対応の抽象化タグを補充する。
図４はその一部を示すものであり、例えば、「東洋自動車」に関しては企業名辞書に登録例が存在していたため、形態素解析処理部12は「東洋（名詞-一般）」と「自動車（名詞-一般）」の形態素を結合した上で、＜企業名＞という抽象化タグを品詞項目に追記する。
【００３２】
「パイロン」に関しても企業名辞書に登録例が存在していたため、形態素解析処理部12は「企業名」という抽象化タグを品詞項目に追記する。
また、「変速機」に関しては活動対象物辞書に生産対象物、販売対象物、開発対象物として登録されていたため、形態素解析処理部12は「変速（名詞-サ変接続）」と「機（名詞-接尾）」の形態素を結合した上で、＜生産対象物＞＜販売対象物＞＜開発対象物＞という抽象化タグを品詞項目に追記する。なお、＜生産対象物＞等の代わりに、上位概念である＜活動対象物＞の抽象化タグを用いることも当然に可能である。
また、「広州市」に関しては地域名辞書に登録例が存在していたため、「広州（名詞-固有名詞）」と「市（名詞-接尾）」の形態素を結合した上で、＜地域＞という抽象化タグが品詞項目に追記される。
さらに、「生産」に関しては企業活動辞書に登録例が存在していたため、＜生産活動＞という抽象化タグが品詞項目に追記される。＜生産活動＞の代わりに、上位概念である＜企業活動＞の抽象化タグを用いることも当然に可能である。
【００３３】

つぎに、複合語解析処理部14が起動し、複合語解析ルール記憶部28に格納された複合語解析ルールを参照することにより、形態素解析処理部12によって形態素単位に分解された文の中から複合語を認定する。

この複合語解析ルールは、図５(a)に示すように、品詞連結パターンと品詞決定基準のデータ項目を備えており、複合語解析処理部14は、文中において品詞連結パターンに合致する形態素の並びを発見すると、これらの形態素を複合語として連結すると共に、対応の品詞決定基準に従い、当該複合語の品詞を同定する。

【００３４】

例えば、図５(b)に示すように、文中に「自然（名詞-形容動詞語幹）」「言語（名詞-一般）」「処理（名詞-サ変接続）」の３つの形態素が連続していた場合、複合語解析処理部14はそれぞれの品詞の連結パターンが複合語解析ルールの(1)にマッチするため「自然言語処理」の複合語と認定した後、(1)の品詞決定基準に基づいてその品詞を「名詞-一般」と認定する。

【００３５】

また、図５(c)に示すように、文中に「高級（名詞-形容動詞語幹）」「化粧品（名詞-一般）」の２つの形態素が連続していた場合、複合語解析処理部14はそれぞれの品詞の連結パターンが複合語解析ルールの(2)にマッチするため「高級化粧品」の複合語と認定した後、(2)の品詞決定基準に基づいてその品詞を「名詞-一般」と認定する。

【００３６】

さらに、図５(d)に示すように、文中に「生産（名詞-サ変接続）」「量（名詞-接尾）」の２つの形態素が連続していた場合、複合語解析処理部14はそれぞれの品詞の連結パターンが複合語解析ルールの(3)にマッチするため「生産量」の複合語と認定した後、(3)の品詞決定基準に基づいてその品詞を「名詞-一般」と認定する。

【００３７】

つぎに、抽象化処理部16が起動し、文中の形態素に対して企業名、生産活動、販売活動、生産対象物等の抽象化タグを関連付ける。

上記のように、先に形態素解析処理部12が辞書記憶部26を参照し、辞書に収録された企業名や企業活動、生産対象物等に対して該当の抽象化タグが付与されているが、辞書の収録語数には自ずと限界があり、辞書ベースでの抽象化処理だけでは漏れが生じる可能性がある。

このため、抽象化処理部20は正規表現ルールによる抽象化処理を実行し、辞書に収録されていない企業名や活動対象物について、対応の抽象化タグを関連付ける機能を備えている。

【００３８】

例えば、「新製品であるABCを～」という表現が文中に存在した場合、「ABC」の部分を「生産対象物」と認定し、「ABCを」の文節に「生産対象物」の抽象化タグを割り当てることを意味する。あるいは、「小売り大手の米AAAマートは、～」という表現が文中に存在した場合に、「AAAマート」の部分を「企業名」と認定し、「AAAマートは」の文節に「企業名」の抽象化タグを割り当てることが該当する。

このため、抽象化ルール記憶部28には、予め多数の抽象化ルールが格納されている。

【００３９】

図６(a)は抽象化ルールの一例を示すものであり、「＜company_size＞の＜country＞（＜feature:名詞＞＋）」は、「company_size（企業規模を表す文字列）」＋「の」＋「country（国を表す文字列）」の直後に続く名詞を企業名と認定することが定義されている。また、「company_size」のエイリアス表現（別名）として、「首位、大手、中堅」が定義されており、「country_size」のエイリアス表現として、「米、英、欧州」が定義されている。

【００４０】

ここに、図６(b)に示すように、「小売大手の米AAAマートは、人員削減計画を発表した。」という文が与えられた場合、抽象化処理部20はこれを図６(c)に示すように名詞単位

のOR表現に置き換え、ルールにマッチする「小売り大手の米AAAマート」を抽出した後、正規表現の「後方参照」を用いて「AAAマート」を取り出し、企業名と認定する。

【００４１】

つぎに、照応解析処理部18が起動し、照応解析ルール記憶部32に格納された照応解析ルールを参照することにより、文中の照応詞（代名詞等）に対して先行詞を補充する。

この照応語解析ルールは、図７(a)に示すように、照応詞と先行詞決定基準のデータ項目を備えており、照応解析処理部18は、定義された照応詞を文中において発見すると、対応の先行詞決定基準に従い、当該照応詞の先行詞を同定する。

【００４２】

例えば、図７(b)に示すように、「同社は　同製品を　１４日より　販売する。」という文が存在した場合、まず照応解析処理部18は「同社」が照応解析ルール(2)の先行詞に該当することを検知し、その先行詞決定基準に従い直近の＜企業名＞タグが付された「Ｂ社」を先行詞と認定し、文中の「同社」と置き換える。

つぎに照応解析処理部18は、文中の「同製品」が照応解析ルール(3)の先行詞に該当することを検知し、その先行詞決定基準に従い直近の＜生産対象物＞タグが付された「新型パソコン」を先行詞と認定し、文中の「同製品」と置き換える。

【００４３】

つぎに、文タイプ判定処理部19が起動し、各文の中で「タイトル文」に該当するものに対しては、タイトル文であることを示す識別情報を付与する。与えられた文がタイトル文であるのか、通常の文（本文）であるのかについては、テキストファイルの収集元であるWebファイルに記述されたタグ情報によって判定される。タイトル文の具体例については、後述する。

また文タイプ判定処理部19は、各文の中で「倒置表現」を含むものに対して、倒置表現文であることを示す識別情報を付与する。倒置表現文の認定方法及び具体例については、後述する。

「タイトル文」または「倒置表現文」の識別情報が付与された文については、次段における格スロット充填処理において、これらの識別情報が付与されていない通常の文とは異なる扱いを受けることとなる。

【００４４】

つぎに、格スロット充填処理部20が起動し、メモリ上に設定された格スロットに対する語（形態素または複合語）の充填処理を実行する。

図８は、格スロットの一例を示すものであり、「助詞」と「対応語」の項目を備えている。また、助詞の項目には、予め（は）、（が）、（を）、（に）…等の必要な助詞（係助詞、格助詞）が設定されている。

【００４５】

ここで図９に示すように、「ソミーは2010年より太陽電池セルを販売する。」という文が与えられた場合、格スロット充填処理部20は格スロットの該当箇所に語を文頭から順に充填する。例えば、「ソミーは」の文節は係助詞の「は」を含んでいるため、同文節内の自立語である「ソミー」が（は）の対応語格納欄に充填される。同様に、「2010年より」の文節は格助詞の「より」を含んでいるため、その直前の語である「2010年」が（より）の対応語格納欄に充填される。同様に「太陽電池セルを」の文節は格助詞の「を」を含んでいるため、その直前の語である「太陽電池セル」が（を）の対応語格納欄に充填される。なお、（は）、（を）、（より）以外の助詞の対応語格納欄については、空欄のまま残される。

【００４６】

つぎに格スロット充填処理部20は、当該格スロットに対して、文の述語である「販売」を関連付ける。

一般的に「述語」といえば、主語の動作や状態、性質などを叙述する動詞、形容詞、名詞＋判定詞を意味するが、格スロット充填処理部20が文中から抽出する「述語」は、最終的な抽出対象である企業情報の「述語」となるべき語であり、具体的には企業活動を示す

＜生産活動＞、＜販売活動＞、＜開発活動＞等の抽象化タグが付された語が該当する。

【００４７】

つぎに、図１０に示すように、「ソミーは2008年より太陽電池技術の研究開発に着手しており、約２年で製品化へ踏み出す。」という文が与えられた場合、格スロット充填処理部20は上記と同様、文頭から順に語の格スロットへの充填処理を実行する。

【００４８】

この際、図８に示した空の格スロットが用いられるのではなく、対応語の充填が完了した直前の格スロットがコピーされ、つぎの文の語によって該当欄に上書充填されるのが原則であるが、つぎの文において（は）格の対応語格納欄に有意思主体を表す語（企業名や人名等の抽象化タグが付与された語）が上書された場合、格スロット充填処理部20は話題が転換されたものと判断し、対応語の継承をキャンセルする。

【００４９】

具体的には、図１０(a)に示すように、前の文から一旦継承した格スロットに対して、後の文の「ソミー（企業名＝有意思主体）」によって（は）格の対応語格納欄が上書された結果、（を）の「太陽電池セル」及び（より）の「2010年」が削除されると同時に、（に）に対して「研究開発」が、（より）に対して「2008年」が、（の）に対して「太陽電池技術」が充填される。

この新たな格スロットに対しては、格スロット充填処理部20によって「研究開発」の述語が関連付けられる。

【００５０】

つぎに格スロット充填処理部20は、図１０(b)に示すように、文の残りの部分である「約２年で製品化へ踏み出す。」の格スロットへの充填処理に移行する。この場合は、文中に（は）を含む文節自体が存在せず、前の文から継承した格スロットの（は）に対する上書充填が生じないため、継承した格スロットの対応語のクリアは行われない。

したがって、（は）には「ソミー」が、（に）には「研究開発」が、（より）には「2008年」が、（の）には「太陽電池技術」がそのまま保持されると共に、（へ）には「製品化」が、（で）には「約２年」が新たに充填される。

この新たな格スロットに対しては、格スロット充填処理部20によって「製品化」の述語が関連付けられる。

【００５１】

つぎに、図１１に示すように、「競合企業であるハープは、2009年より太陽電池セルの販売を開始している。」という文が与えられた場合、格スロット充填処理部20は前の文の格スロットを一旦継承させるが、（は）の対応語格納欄に有意思主体である「ハープ（＜企業名＞）」が上書充填された時点で、話題の転換が生じたものと判断して他の対応語格納欄に充填された対応語をクリアした後、改めて（を）に「販売」を、（より）に「2009年」を、（の）に「太陽電池セル」を充填する。

この新たな格スロットに対して格スロット充填処理部20は、「販売」の述語を関連付ける。

【００５２】

上記した格スロット充填処理は、通常の自然文を対象とした場合の例であるが、与えられた文がタイトル文の場合、助詞が省略されていることが多いため、格スロット充填処理部20は省略された助詞を推定した上で、格スロットに対する語の充填処理を実行する。

【００５３】

例えば、図１２に示すように、「ソミー、次世代太陽電池部品販売」のタイトル文が与えられた場合、格スロット充填処理部20は企業名（有意思主体）である「ソミー」については（は）または（が）の助詞が省略されているものと推定し、格スロットの（は）及び（が）に「ソミー」を充填する。

つぎに格スロット充填処理部20は、「次世代太陽電池部品」について、＜販売対象物＞の抽象化タグが付与されていることから、目的語を表す助詞である（を）が省略されているものと推定し、格スロットの（を）に「次世代太陽電池部品」を充填する。

つぎに格スロット充填処理部20は、「販売」について、企業活動を表す＜販売活動＞の抽象化タグが付与されているため述語であると認定し、格スロットに述語として「販売」を関連付ける。

【００５４】

与えられた文が倒置表現を含む場合にも、格スロット充填処理部20は特別な充填処理を実行する。例えば、図１３に示すように、「Ａ社が独自に開発した高速無線通信技術を採用した。」という文の場合、前半の「Ａ社（＜企業名＞）が独自に開発（＜開発活動＞）した」の部分は目的語を有さない不完全な文となってしまう。

【００５５】

このような文に対しては、事前に文タイプ判定処理部19が述語である「開発」の前に「ヲ格（目的格）」が存在するか否かをチェックし、存在しない場合には述語について倒置表現が用いられているものと判断し、「倒置表現文」の識別情報を付与している。

【００５６】

そこで格スロット充填処理部20は、当該述語（開発）に後続する名詞を当該文のヲ格と認定し、「高速無線通信技術」を格スロットの（を）に充填する。

また、格スロット充填処理部20は、この格スロットに対して「開発」の述語を関連付ける。

【００５７】

格スロット充填処理部20によって必要な語が充填された格スロットに対しては、情報抽出処理部22が抽出フレーム定義記憶部34に格納された抽出フレーム定義を適用することにより、主語、述語、目的語の３つの要素を備えた情報（所謂トリプル）を抽出する。

【００５８】

図１４は、この抽出フレーム定義の適用例を示すものであり、抽出フレーム定義50には、以下の(1)～(3)の条件を全て満たしている場合に、当該格スロットからトリプルを抽出すべきことが規定されている。

(1)格スロットの（が）または（は）に＜企業名＞の抽象化タグが付与された語が充填されていること。

(2)格スロットの述語として＜販売活動＞の抽象化タグが付与された語が関連付けられていること。

(3)格スロットの（を）に＜生産対象物＞の抽象化タグが付与された語が充填されていること。

【００５９】

情報抽出処理部22は、格スロット充填処理部20から渡された充填済みの格スロット52に対して、上記抽出フレームを定義を当てはめ、上記の(1)～(3)の条件に合致する場合には、当該格スロットの（が）または（は）に充填された語を「主語」とし、（を）に充填された語を「目的語」とし、当該格スロットに関連付けられた述語を「述語」とするトリプル54を生成する。

このトリプルは、情報抽出処理部22により、RDF（Resource Description Framework）形式の企業情報としてセマンティックＤＢ24に格納される。

【００６０】

抽出フレーム定義記憶部34には、目的に応じて多数の抽出フレーム定義が格納される。

例えば、企業の製品開発情報を収集したい場合には、「主語：＜企業名＞（が｜は）」、「述語：＜開発活動＞」、「目的語：＜開発対象物＞」の抽出フレーム定義を用意しておけばよい。

【００６１】

情報抽出処理部22は、上記のトリプル抽出処理に際し、抽出制限ルール記憶部36に格納された抽出制限ルールを参照し、適合する格スロットに対しては必要な処理を実行する。

ここで、抽出制限ルールとしては、「削除対象定義」と、「除外対象定義」が規定されている。以下、個別に説明する。

【００６２】

まず、「削除対象定義」とは、抽出対象となる語から不要な形容詞を除去するためのルールを規定するものである。

その一例として、「^(?:新型|次期|次世代)(.+)/ -> $1」という削除対象定義は、「新型」「次期」「次世代」という形容詞の削除を規定しているため、格スロットの（を）に「新型パソコン」や「次期ハイブリッド車」が充填されていた場合には、情報抽出処理部22によって「パソコン」や「ハイブリッド車」に整形された上で、トリプルの一部として抽出される。

【００６３】

つぎに、「除外対象定義」とは、格スロットに目的語として格納されている語が、情報として役に立たない抽象的な表現のものである場合に、これを抽出対象から除外すべく規定されているものであり、例えば「新製品」や「戦略車」などが該当する。

情報抽出処理部22は、格スロット充填処理部20から渡された格スロットの（を）の対応語をチェックし、そこに「新製品」や「戦略車」などが充填されていた場合、当該格スロットからのトリプル抽出処理をキャンセルする。

【００６４】

上記においては、「主語－述語－目的語」を備えたトリプルをテキストから抽出する例を示したが、オプションとして場所や時間を示す情報要素を抽出することもできる。

例えば、抽出フレーム定義中に「オプション：＜地域＞で」の条件を加えておけば、格スロットの（で）の対応語格納欄に＜地域＞の抽象化タグが付与された語（例えば「広州市」）が充填されている場合、情報抽出処理部22は主語、述語、目的語に該当する文字列と共に、この地域を表す文字列を抽出する。

あるいは、抽出フレーム定義中に「オプション：＜時間＞より」の条件を加えておけば、格スロットの（より）の対応語格納欄に＜時間＞の抽象化タグが付与された語（例えば「2010年」）が充填されている場合、情報抽出処理部22は主語、述語、目的語に該当する文字列と共に、この時間を表す文字列を抽出する。

【００６５】

上記のように、セマンティック企業ＤＢ24に蓄積された企業情報は、企業名、企業活動（生産活動、販売活動等）、活動対象物（生産対象物等）の明確な意味的構造を備えているため、これを検索用データベースとして利用することにより、極めて効率的な企業情報の検索が可能となる。

【００６６】

例えば、クライアント端末42から「小麦粉 AND 販売」という検索条件が送信された場合、検索サーバ38はセマンティック企業ＤＢ24から小麦粉を販売対象としている企業情報を抽出し、企業名のリストをクライアント端末42に送信することができる。

あるいは、クライアント端末42から「東北地方 AND 工場」という検索条件が送信された場合、検索サーバ38はセマンティック企業ＤＢ24を検索し、東北地方で工場に関する何らかの活動（例えば保有、建設、賃貸、閉鎖等）を行っている企業名のリストをクライアント端末42に送信する。より絞り込まれた情報を希望するユーザは、「東北地方 AND 工場 AND 保有」のように検索条件を変更すればよい。

また、「東洋自動車 AND 生産 AND 中国」の検索条件が送信された場合、検索サーバ38は東洋自動車が中国で生産している対象物のリストを生成し、クライアント端末42に送信することができる。

【００６７】

従来の構文解析技術を用いた情報抽出方式の場合、抽出すべき文節間の係り受け構造を構文パターンとして定義しておく必要があるが、このように文の構造に依存する構文パターンを用いる方式では、語の順番が入れ替わっただけでも対象となる情報の抽出が不可能となるため、文のあらゆるバリエーションを想定して構文パターンを準備する必要があった。

【００６８】

これに対し、この情報抽出システム10によれば、上記のように予め必要な助詞の種類が

設定された定型的な格スロットと、抽出すべき情報の種類を規定する抽出フレーム定義を用意しておき、文中に対応の助詞が存在する場合にはその直前の自立語を当該助詞の対応語格納欄に充填すると共に、この充填済みの格スロットに抽出フレーム定義を適用することにより、語順にかかわらず自然文から「主語－述語－目的語」のように構造化された情報を確実に抽出することができる。また、語順に拘束されないため、抽出フレーム定義のバリエーションを抑制することができる。

しかも、原則として前の文の格スロットがつぎの文に継承される仕組みを備えているため、後続の文中に主語や目的語の省略が存在したとしても、前の文の主語や目的語で容易に補うことができる。

【００６９】

上記においては、格スロット充填処理部20から渡された充填済みの格スロットに対して、情報抽出処理部22がトリプル抽出処理を直ちに実行する例を説明したが、この発明はこれに限定されるものではない。

すなわち、格スロット充填処理部20によって必要な語の充填が完了した格スロットを充填済み格スロット記憶部（図示省略）に蓄積しておき、これに対し情報抽出処理部22が多種多様な抽出フレーム定義を順次適用することにより、各種情報をまとめて抽出するように構成してもよい。

【００７０】

上記においては、テキスト文から「企業名（主語）→企業活動（述語）→活動対象物（目的語）」のトリプル構造を備えた企業情報を抽出する例を示したが、この発明はこれに限定されるものではない。

例えば、企業名辞書の代わりに人名辞書を、企業活動辞書の代わりに人間活動辞書を、企業の活動対象物辞書の代わりに人間の活動対象物辞書を用意し、人間の活動を抽出するための抽出フレーム定義を準備しておくことにより、「人名（主語）→人間活動（述語）→活動対象物（目的語）」のトリプル構造を備えた人物情報（芸能人情報等）を抽出することも可能となる。

【図面の簡単な説明】

【００７１】

【図１】この発明に係る情報抽出システムの機能構成を示すブロック図である。
【図２】企業活動辞書の登録内容を例示する図表である。
【図３】活動対象物辞書の登録内容を例示する図表である。
【図４】形態素解析の結果を示す図表である。
【図５】複合語解析ルール及びその適用事例を示す説明図である。
【図６】抽象化ルールによる抽象化処理を示す説明図である。
【図７】照応解析ルール及びその適用事例を示す説明図である。
【図８】格スロットの構成を示す概念図である。
【図９】格スロットに対する語の充填例を示す説明図である。
【図１０】格スロットに対する語の充填例を示す説明図である。
【図１１】格スロットに対する語の充填例を示す説明図である。
【図１２】格スロットに対する語の充填例を示す説明図である。
【図１３】格スロットに対する語の充填例を示す説明図である。
【図１４】抽出フレーム定義及びその適用事例を示す説明図である。
【図１５】構文解析技術を用いた従来の情報抽出方法を示す説明図である。
【図１６】構文解析技術を用いた従来の情報抽出方法を示す説明図である。
【図１７】構文解析技術を用いた従来の情報抽出方法を示す説明図である。
【図１８】構文解析技術を用いた従来の情報抽出方法を示す説明図である。

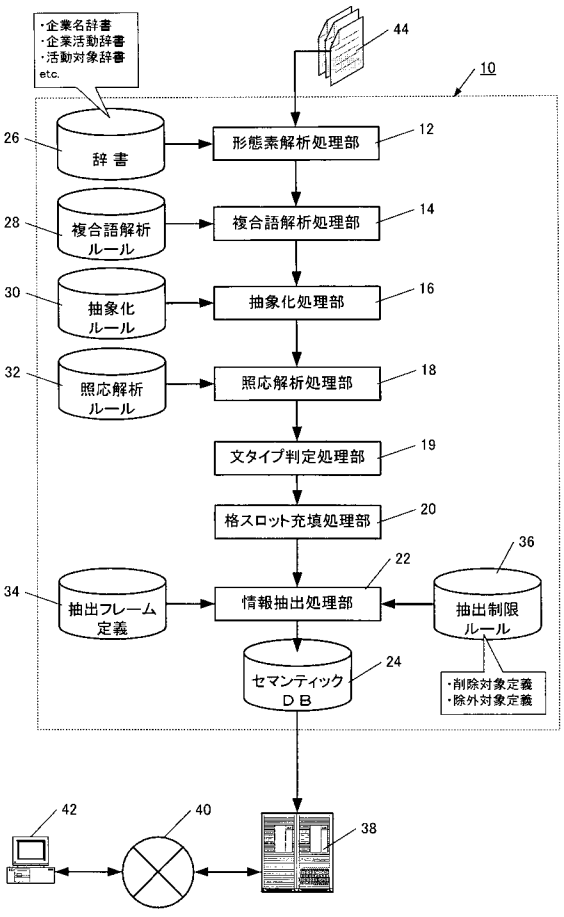
【符号の説明】

【００７２】

10　情報抽出システム
12　形態素解析処理部

14　複合語解析処理部
16　抽象化処理部
18　照応解析処理部
19　文タイプ判定処理部
20　格スロット充填処理部
22　情報抽出処理部
24　セマンティック企業ＤＢ
26　辞書記憶部
28　複合語解析ルール記憶部
30　抽象化ルール記憶部
32　照応解析ルール記憶部
34　抽出フレーム定義記憶部
36　抽出制限ルール記憶部
38　検索サーバ
40　通信ネットワーク
42　クライアント端末
44　テキストデータ

【図１】



【図２】

抽象化文字列	表現文字列
生産活動	生産
	製造
	加工
	組立
⋮	
販売活動	販売
	発売
	売り出す
⋮	
開発活動	開発
	研究
	研究開発
⋮	

【図３】

抽象化文字列	表現文字列
生産対象物	液晶
	液晶テレビ
	液晶パネル
	液晶モニター
⋮	
販売対象物	液晶
	液晶テレビ
	液晶パネル
	液晶モニター
⋮	
開発対象物	液晶
	液晶テレビ
	液晶パネル
	液晶モニター
⋮	

【図４】

形態素	品詞
東洋自動車	名詞，＜企業名＞
子会社	名詞−一般
の	助詞−連体化
変速機	名詞，＜生産対象物＞＜販売対象物＞＜開発対象物＞
メーカー	名詞−一般
パイロン	名詞，＜企業名＞
⋮	
広州市	名詞，＜地域＞
で	助詞−格助詞
生産	名詞，＜生産活動＞
する	動詞−自立
と	助詞−格助詞
発表	名詞−サ変接続
し	動詞−自立
た	助動詞
。	記号−句点

【図５】

(a)

（複合語解析ルール）

NO.	品詞連結パターン	品詞決定基準
(1)	[名詞−形容動詞語幹]+[名詞−一般]+[名詞−サ変接続]	後続形態素が[サ変・スル]→[名詞−サ変接続]、それ以外→[名詞−一般]
(2)	[名詞−形容動詞語幹]+[名詞−一般]	末尾の形態素の品詞
(3)	[名詞−サ変接続]+[名詞−接尾]	[名詞−一般]
⋮		

(b) 自然[名詞−形容動詞語幹]／言語[名詞−一般]／処理[名詞−サ変接続]

⇩ ※ (1)を適用

自然言語処理[名詞−一般]

(c) 高級[名詞−形容動詞語幹]／化粧品[名詞−一般]

⇩ ※ (2)を適用

高級化粧品[名詞−一般]

(d) 生産[名詞−サ変接続]／量[名詞−接尾]

⇩ ※ (3)を適用

生産量[名詞−一般]

【図６】

(a)

企業名の抽象化ルール

<company_size>の<country>(<feature:名詞>+)

<company_size>
-首位
-大手
-中堅

<country>
-米
-英
-欧州

エイリアス表現

(b)

処理対象文

「小売大手の米AAAマートは、人員削減計画を発表した。」

(c)

正規表現

(:?首位|大手|中堅)の(:?米|英|欧州) ((:?小売|大手|米|AAAマート|人員|削減|計画|発表)+)

⇩

「小売り大手の米AAAマート」

⇩

「AAAマート」:企業名

【図７】

(a)

(照応解析ルール)

NO.	照応詞	先行詞決定基準
(1)	「彼」	直近の[名詞-固有名詞-人名」の形態素を先行詞とする
(2)	^同社$	直近の＜企業名＞を先行詞とする
(3)	^(?:同|新)製品$	直近の＜生産対象物＞を先行詞とする

⋮

(b)

B社は 新型パソコンを 発表した。 同社は 同製品を 14日より 販売する。
＜企業名＞ ＜生産対象物＞

(照 応)

B社は 新型パソコンを 発表した。 B社は 新型パソコンを 14日より 販売する。
＜企業名＞ ＜生産対象物＞ ＜企業名＞ ＜生産対象物＞

【図８】

[格スロット]

助 詞	対応語
(は)	
(が)	
(を)	
(に)	
(から)	
(まで)	
(より)	
(へ)	
(と)	
(や)	
(の)	
(で)	

⋮

【図９】

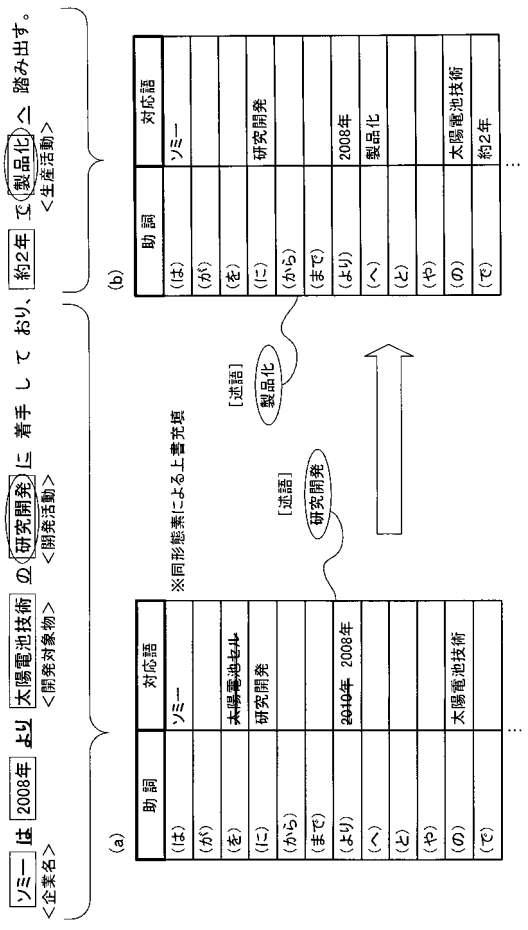
ソミー は 2010年 より 太陽電池セル を 販売 する。
＜企業名＞ ＜販売対象物＞ ＜販売活動＞

助 詞	対応語
(は)	ソミー
(が)	
(を)	太陽電池セル
(に)	
(から)	
(まで)	
(より)	2010年
(へ)	
(と)	
(や)	
(の)	
(で)	

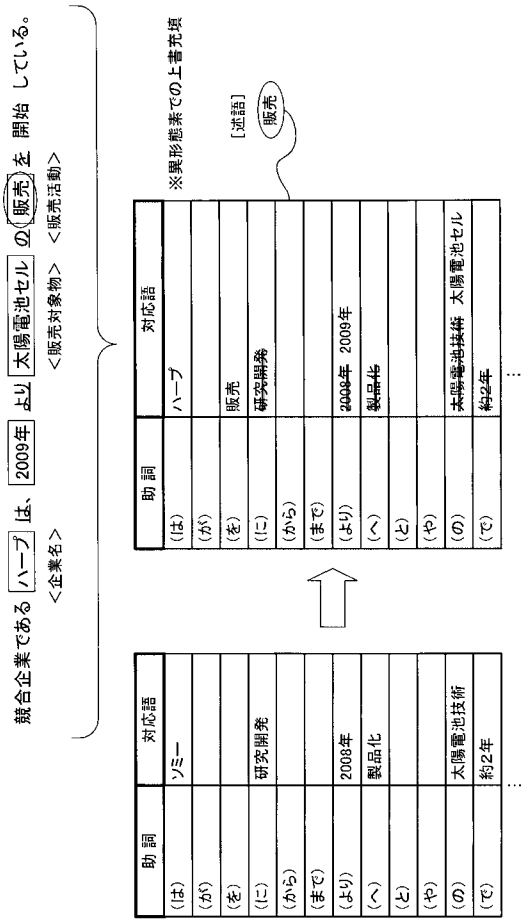
[述語] 販売

⋮

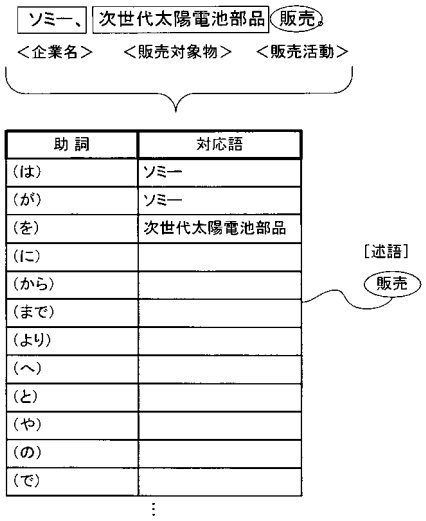
【図１０】



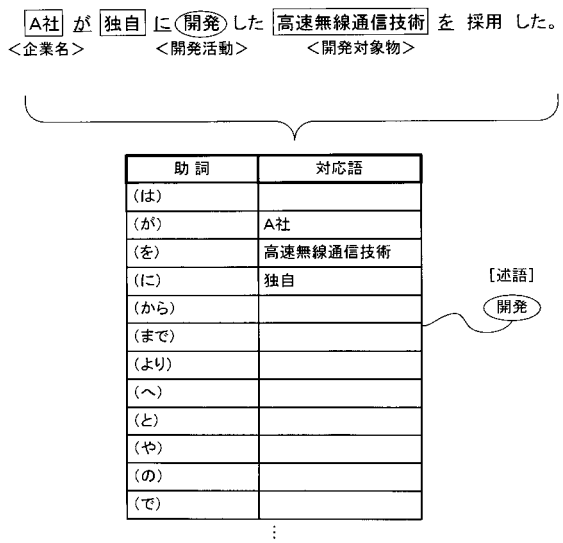
【図１１】



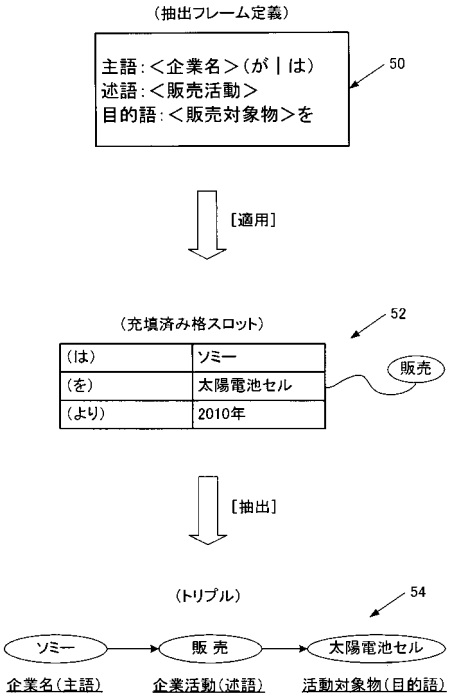
【図１２】



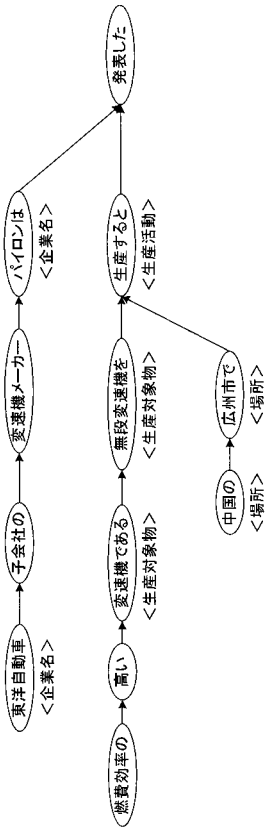
【図１３】



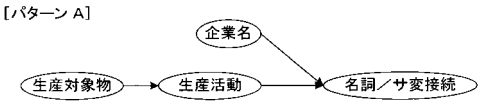
【図１４】



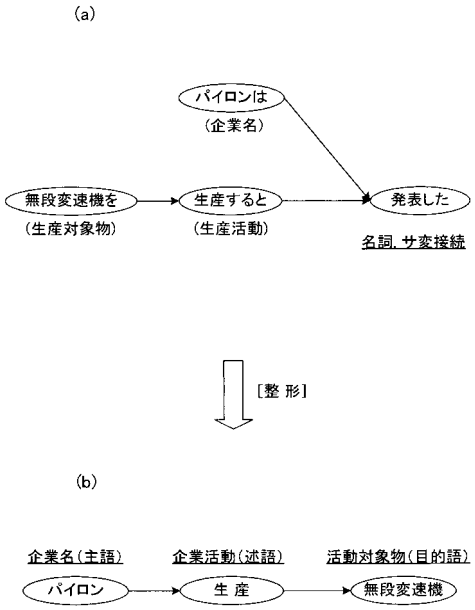
【図１５】



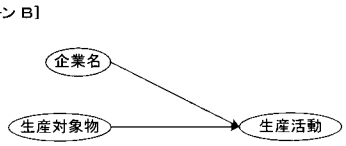
【図１６】



【図１７】



【図１８】



【手続補正書】
【提出日】平成22年10月8日(2010.10.8)
【手続補正１】
【補正対象書類名】特許請求の範囲
【補正対象項目名】全文
【補正方法】変更
【補正の内容】
【特許請求の範囲】
【請求項１】
活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、
上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、
上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、
テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段と、
上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には、当該形態素に対してその種類を示す抽象化文字列を関連付ける手段と、
少なくとも主語に付属する助詞毎及び目的語に付属する助詞毎に対応語の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化文字列が関連付けられている形態素を述語として当該格スロットに関連付ける格スロット充填手段と、
抽出すべき主語の抽象化文字列及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化文字列を特定する条件と、抽出すべき目的語の抽象化文字列及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段と、
対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する語を抽出する情報抽出手段とを備えた情報抽出システムであって、
上記格スロット充填手段が、以下の処理を実行することを特徴とする情報抽出システム。
(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。
(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。
(3) 文の主語に付属する助詞の対応語格納欄が後続の述語に係る企業名や人名を示す抽象化文字列が関連付けられた語によって上書充填された場合には、先行する述語に関して充填された対応語を削除する。
【請求項２】
上記の文の中でタイトルに該当する文に対して、タイトル文であることを示す識別情報を予め付与する手段を備え、
この識別情報が付与されたタイトル文に対して、上記格スロット充填手段が以下の処理を実行することを特徴とする請求項１に記載の情報抽出システム。
(1)タイトル文中の主語となるべき種類の抽象化文字列が関連付けられた語については、助詞の有無を問わず主語に付属する助詞の対応語格納欄に充填する。
(2)タイトル文中の目的語となるべき種類の抽象化文字列が関連付けられた語については、助詞の有無を問わず目的語に付属する助詞の対応語格納欄に充填する。
【請求項３】
文中における述語の前に目的語に付属する助詞が存在しない場合に、倒置表現文であることを示す識別情報を予め付与する手段を備え、
この識別情報が付与された倒置表現文に対して、上記格スロット充填手段が当該述語に

後続する名詞を文の目的語と認定し、格スロットの目的語に付属する助詞の対応語格納欄に当該名詞を充填することを特徴とする請求項１または２に記載の情報抽出システム。
【請求項４】
複合語となるべき複数の品詞の連結パターン毎に、当該複合語の品詞を決定するための基準が規定された複合語解析ルールを格納しておく複合語解析ルール記憶手段と、
この複合語解析ルールを参照し、文中に複合語解析ルールに規定された品詞の連結パターンに該当する形態素の組合せが存在している場合には、これらの形態素を複合語と認定する複合語解析手段とを備え、
上記の格スロット充填手段は、複合語と認定された形態素の組合せについては、複合語単位で格スロットへの充填処理を実行することを特徴とする請求項１～３の何れかに記載の情報抽出システム。
【請求項５】
形態素の種類を推定するための抽象化ルールを格納しておく抽象化ルール記憶手段と、
上記の抽象化ルールを文に対して適用し、当該抽象化ルールにマッチする形態素に対してその種類を示す抽象化文字列を関連付ける手段と、
を備えたことを特徴とする請求項１～４の何れかに記載の情報抽出システム。
【請求項６】
照応詞毎に、その先行詞を決定するための基準を定めた照応解析ルールを格納しておく照応解析ルール記憶手段と、
この照応解析ルールを参照し、文中に存する照応詞に対して、対応の先行詞を決定すると共に、この先行詞によって照応詞を置き換える照応解析手段とを備えたことを特徴とする請求項１～５の何れかに記載の情報抽出システム。
【請求項７】
コンピュータを、
活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、
上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、
上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、
テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段、
上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には、当該形態素に対してその種類を示す抽象化文字列を関連付ける手段、
少なくとも主語に付属する助詞毎及び目的語に付属する助詞毎に対応語の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化文字列が関連付けられている形態素を述語として当該格スロットに関連付ける格スロット充填手段、
抽出すべき主語の抽象化文字列及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化文字列を特定する条件と、抽出すべき目的語の抽象化文字列及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段、
対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する語を抽出する情報抽出手段として機能させる情報抽出プログラムであって、
上記格スロット充填手段が、以下の処理を実行することを特徴とする情報抽出プログラム。
(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。
(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。
(3) 文の主語に付属する助詞の対応語格納欄が後続の述語に係る企業名や人名を示す抽象化文字列が関連付けられた語によって上書充填された場合には、先行する述語に関して

充填された対応語を削除する。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】００１０
【補正方法】変更
【補正の内容】
【００１０】
上記の目的を達成するため、請求項１に記載した情報抽出システムは、活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書と、テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段と、上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には、当該形態素に対してその種類を示す抽象化文字列を関連付ける手段と、少なくとも主語に付属する助詞（「は」、「が」）毎及び目的語に付属する助詞（「を」、「に」等）毎に対応語（自立語）の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化文字列が関連付けられている形態素を述語として当該格スロットに関連付ける格スロット充填手段と、抽出すべき主語の抽象化文字列及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化文字列を特定する条件と、抽出すべき目的語の抽象化文字列及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段と、対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する語を抽出する情報抽出手段とを備えた情報抽出システムであって、上記格スロット充填手段が、以下の処理を実行することを特徴としている。
(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。
(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。
(3) 文の主語に付属する助詞の対応語格納欄が後続の述語に係る企業名や人名を示す抽象化文字列が関連付けられた語によって上書充填された場合には、先行する述語に関して充填された対応語を削除する。
【手続補正３】
【補正対象書類名】明細書
【補正対象項目名】００１１
【補正方法】変更
【補正の内容】
【００１１】
請求項２に記載した情報抽出システムは、請求項１のシステムであって、さらに、上記の文の中でタイトルに該当する文に対して、タイトル文であることを示す識別情報を予め付与する手段を備え、この識別情報が付与されたタイトル文に対して、上記格スロット充填手段が以下の処理を実行することを特徴としている。
(1)タイトル文中の主語となるべき種類の抽象化文字列が関連付けられた語については、助詞の有無を問わず主語に付属する助詞（「は」、「が」）の対応語格納欄に充填する。
(2)タイトル文中の目的語となるべき種類の抽象化文字列が関連付けられた語については、助詞の有無を問わず目的語に付属する助詞（「を」、「に」等）の対応語格納欄に充填する。
【手続補正４】
【補正対象書類名】明細書
【補正対象項目名】００１２

【補正方法】変更
【補正の内容】
【００１２】
請求項３に記載した情報抽出システムは、請求項１または２のシステムであって、さらに、文中における述語の前に目的語に付属する助詞（例えば「を」）が存在しない場合に、倒置表現文であることを示す識別情報を予め付与する手段を備え、この識別情報が付与された倒置表現文に対して、上記格スロット充填手段が当該述語に後続する名詞を当該述語の目的語と認定し、格スロットの目的語に付属する助詞の対応語格納欄に当該名詞を充填することを特徴としている。
【手続補正５】
【補正対象書類名】明細書
【補正対象項目名】００１４
【補正方法】変更
【補正の内容】
【００１４】
請求項５に記載した情報抽出システムは、請求項１～４のシステムであって、さらに、形態素の種類を推定するための抽象化ルールを格納しておく抽象化ルール記憶手段と、上記の抽象化ルールを文に対して適用し、当該抽象化ルールにマッチする形態素に対してその種類を示す抽象化文字列を関連付ける手段とを備えたことを特徴としている。
【手続補正６】
【補正対象書類名】明細書
【補正対象項目名】００１６
【補正方法】変更
【補正の内容】
【００１６】
請求項７に記載した情報抽出プログラムは、コンピュータを、活動主体となる具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、上記活動主体の活動内容を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、上記活動主体の活動対象物を示す具体的な表現文字列と、その種類を示す抽象化文字列との対応関係を登録した辞書、テキストデータ中の文を形態素単位に分解し、各形態素の品詞を同定する手段、上記の各辞書を参照し、文中において各辞書に収録されている形態素が存する場合には当該形態素に対してその種類を示す抽象化文字列を関連付ける手段、少なくとも主語に付属する助詞毎及び目的語に付属する助詞毎に対応語の格納欄が設けられた格スロットに、文中の対応語を充填すると共に、活動内容を示す抽象化文字列が関連付けられている形態素を述語として当該格スロットに関連付ける格スロット充填手段、抽出すべき主語の抽象化文字列及び当該主語に付属する助詞を特定する条件と、抽出すべき述語の抽象化文字列を特定する条件と、抽出すべき目的語の抽象化文字列及び当該目的語に付属する助詞を特定する条件が少なくとも規定された抽出フレーム定義を、複数格納しておく抽出フレーム定義記憶手段、対応語充填済みの上記格スロットに上記抽出フレーム定義を適用することにより、少なくとも文の主語、述語、目的語に該当する語を抽出する情報抽出手段として機能させる情報抽出プログラムであって、上記格スロット充填手段が、以下の処理を実行することを特徴としている。
(1) 先行する述語に関して対応語の充填が完了した格スロットを、後続の述語について継承させる。
(2) 後続の述語に係る対応語を上記格スロットの対応語格納欄に上書充填する。
(3) 文の主語に付属する助詞の対応語格納欄が後続の述語に係る企業名や人名を示す抽象化文字列が関連付けられた語によって上書充填された場合には、先行する述語に関して充填された対応語を削除する。

フロントページの続き

(72)発明者　岡田　智靖
東京都千代田区丸の内一丁目６番５号　株式会社野村総合研究所内

Ｆターム(参考) 5B075 ND03 NS10
5B091 AA15 AB06 AB19 CA02 CA14 CA21