

(19) 日本国特許庁(JP)

(12) 特 許 公 報(B2)

(11) 特許番号

特許第4934819号
(P4934819)

(45) 発行日 平成24年5月23日(2012.5.23)

(24) 登録日 平成24年3月2日(2012.3.2)

(51) Int.Cl. F 1
G 0 6 F 17/30 (2006.01)
 G 0 6 F 17/30 2 2 0 Z
 G 0 6 F 17/30 1 7 0 A
 G 0 6 F 17/30 2 1 0 A

請求項の数 8 (全 25 頁)

(21) 出願番号 特願2007-94339(P2007-94339)
 (22) 出願日 平成19年3月30日(2007.3.30)
 (65) 公開番号 特開2008-250887(P2008-250887A)
 (43) 公開日 平成20年10月16日(2008.10.16)
 審査請求日 平成21年9月25日(2009.9.25)

(73) 特許権者 504174135
 国立大学法人九州工業大学
 福岡県北九州市戸畑区仙水町1番1号
 (74) 代理人 100099634
 弁理士 平井 安雄
 (72) 発明者 野村 浩郷
 福岡県飯塚市大字川津680-4 九州工
 業大学内
 審査官 鈴木 和樹

最終頁に続く

(54) 【発明の名称】 情報抽出装置、その方法及びプログラム

(57) 【特許請求の範囲】

【請求項1】

記事入力手段と、
 入力された記事情報についてテンプレートを利用して主要な製品情報を抽出するテンプレート抽出手段と、
 入力された記事情報について付与されたタグに基づいてパターンマッチングを行うダグパターンマッチング手段と、
 前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手段と、
 分割された記事の見出しを形態素解析する見出しの形態素解析手段と、
 形態素解析された見出しの文節から助詞を除去する見出し助詞除去手段と、
 前記テンプレートにより抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチング手段と、
 前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出手段と、
 分割された記事の本文を形態素解析する本文の形態素解析手段と、
 形態素解析された本文の文節から助詞を除去する本文助詞除去手段と、
 見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手段と、
 前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手段と、
 見出し特徴情報または本文特徴情報を売り情報として出力する売り情報の出力手段と

10

20

を含む情報抽出装置。

【請求項 2】

記事入力手段と、
 入力された記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出手段と、
 入力された記事情報について付与されたタグに基づいてパターンマッチングを行うダグパターンマッチング手段と、
 前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手段と、
 分割された記事の見出しを形態素解析する見出しの形態素解析手段と、
 形態素解析された見出しの文節から助詞を除去する見出し助詞除去手段と、
 前記係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチング手段と、
 前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出手段と、
 分割された記事の本文を形態素解析する本文の形態素解析手段と、
 形態素解析された本文の文節から助詞を除去する本文助詞除去手段と、
 見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手段と、
 前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手段と、
 見出し特徴情報または本文特徴情報を売り情報として出力する売り情報の出力手段と
 を含む情報抽出装置。

10

20

【請求項 3】

記事入力手段と、
 抽出精度の重み付けの閾値が一定以上のテンプレートを利用して入力された記事情報から主要な製品情報を抽出するテンプレート抽出手段と、
 前記テンプレートにより抽出されなかった記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出手段と、
 入力された記事情報について付与されたタグに基づいてパターンマッチングを行うダグパターンマッチング手段と、
 前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手段と、
 分割された記事の見出しを形態素解析する見出しの形態素解析手段と、
 形態素解析された見出しの文節から助詞を除去する見出し助詞除去手段と、
 前記テンプレートまたは係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチング手段と、
 前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出手段と、
 分割された記事の本文を形態素解析する本文の形態素解析手段と、
 形態素解析された本文の文節から助詞を除去する本文助詞除去手段と、
 見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手段と、
 前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手段と、
 見出し特徴情報または本文特徴情報を売り情報として出力する売り情報の出力手段と
 を含む情報抽出装置。

30

40

【請求項 4】

前記形態素解析された本文の係り受け関係を調べる本文の係り受け関係解析手段と、
 係り受け関係により修飾語句を補足説明情報として抽出する補足説明の抽出手段と、
 補足説明情報を売り情報として出力する売り情報の出力手段と
 を含む請求項 1 ないし請求項 3 のいずれかに記載された情報抽出装置。

【請求項 5】

前記テンプレート抽出手段は、

50

入力された記事を句点ごとに分割する記事句点分割手段と、
記事の1行目に対応するAテンプレート集合とマッチングを行うAテンプレートマッチング手段と、

前記Aテンプレート集合によりマッチングされた製品の特徴情報を抽出するAテンプレート抽出手段と、

抽出された製品の特徴情報について抽出項目ごとに制約をチェックする制約チェック手段と、

情報を抽出することができたテンプレートのIDを記憶するテンプレートID記憶手段と、

記事の2行目以降に対応するBテンプレート集合とマッチングを行うBテンプレートマッチング手段と、

前記Bテンプレート集合によりマッチングされた製品の特徴情報を抽出するBテンプレート抽出手段と、

抽出された製品の特徴情報である抽出解と製品を対応付けるテンプレート製品対応手段と、

を含む請求項1又は3に記載された情報抽出装置。

【請求項6】

前記係り受け抽出手段は、

入力された記事に付与されたタグに基づいてパターンマッチングを行う係り受けタグパターンマッチング手段と、

前記パターンマッチングの結果に基づいて記事を見出しと本文に分割する係り受けタグ分割手段と、

分割された見出しに含まれる特殊記号を分析する見出し分析手段と、

前記見出しに含まれる特殊記号の後方の語句を「販売元」情報として処理する見出し処理手段と、

分割された本文を句点ごとに分割する本文句点分割手段と、

前記本文中に括弧内数値が存在するか否かを判定する括弧内数値判定手段と、

括弧内数値が存在すると判断した場合に、構文解析により文節情報を作成する文節情報作成手段と、

括弧内数値が存在しないと判断した場合に、固定パターンが存在するか否かを判断する固定パターン判定手段と、

固定パターンが存在すると判断した場合に、固定パターンと文節情報から得られる固定パターンの係り受け情報を利用して固定パターンに係る文節情報集合を作成する固定パターン係り受け作成手段と、

前記作成された固定パターンに係る文節情報集合から固定パターン及び各形式について定めた条件に従って文節情報を抽出する抽出手段と、

抽出された文節情報から不要な情報を削除して抽出解を作成する抽出解作成手段と、

抽出解から製品に対する対応や割り当てを行う係り受け対応・割付手段と

を含む請求項2又は3に記載された情報抽出装置。

【請求項7】

記事入力ステップと、

入力された記事情報についてテンプレートを利用して主要な製品情報を抽出するテンプレート抽出ステップと、

前記テンプレートにより抽出されなかった記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出ステップと、

入力された記事情報について付与されたタグに基づいてパターンマッチングを行うタグパターンマッチングステップと、

前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割ステップと、

分割された記事の見出しを形態素解析する見出しの形態素解析ステップと、

10

20

30

40

50

形態素解析された見出しの文節から助詞を除去する見出し助詞除去ステップと、
前記テンプレートまたは係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチングステップと、

前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出ステップと、

分割された記事の本文を形態素解析する本文の形態素解析ステップと、

形態素解析された本文の文節から助詞を除去する本文助詞除去ステップと、

見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチングステップと、

前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出ステップと、

形態素解析された本文の係り受け関係を調べる本文の係り受け関係解析ステップと、

係り受け関係により修飾語句を補足説明情報として抽出する補足説明の抽出ステップと、

見出し特徴情報または本文特徴情報または補足説明情報を売り情報として出力する売り情報の出力ステップと

を含む情報抽出方法。

【請求項 8】

記事入力手順と、

入力された記事情報についてテンプレートを利用して主要な製品情報を抽出するテンプレート抽出手順と、

前記テンプレートにより抽出されなかった記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出手順と、

入力された記事情報について付与されたタグに基づいてパターンマッチングを行うタグパターンマッチング手順と、

前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手順と、

分割された記事の見出しを形態素解析する見出しの形態素解析手順と、

形態素解析された見出しの文節から助詞を除去する見出し助詞除去手順と、

前記テンプレートまたは係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチング手順と、

前記主要な製品情報以外の情報を示す文節として見出しの特徴情報を抽出する見出し特徴情報抽出手順と、

分割された記事の本文を形態素解析する本文の形態素解析手順と、

形態素解析された本文の文節から助詞を除去する本文助詞除去手順と、

見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手順と、

前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手順と、

形態素解析された本文の係り受け関係を調べる本文の係り受け関係解析手順と、

係り受け関係により修飾語句を補足説明情報として抽出する補足説明の抽出手順と、

見出し特徴情報または本文特徴情報または補足説明情報を売り情報として出力する売り情報の出力手順

としてコンピュータを機能させる情報抽出プログラム。

【発明の詳細な説明】

【技術分野】

【0001】

本発明は、文書の中から、あらかじめ定められた種類の情報を自動抽出する装置に関する。

【背景技術】

【0002】

ネットワークの普及により、情報を電子化して管理する機会が増加しつつある現在、そ

10

20

30

40

50

これらの情報の中から必要な情報だけを人間の手作業で取り出してくることは困難な状況になっている。このような状況下において、大量の情報を管理する技術として文章データの中から目的の情報のみを取り出してくる情報抽出の技術が要求されている。また、情報抽出は文章の整理やデータベースの自動的な構築、要約文の生成など応用範囲の広い技術に適用されるため、より高機能・高精度な情報抽出技術が必要となっている。

そこで、情報抽出を利用した文書要約装置が、特開2002-288190号公報に開示されている。

【0003】

背景技術の情報抽出を利用した文書要約装置は、形態素列(1つの形態素からなる列を含む。以下同様。)の入力を受け付ける形態素列入力受付部と、要素として認定されるべき形態素列と、当該要素の属性と、をあらかじめ記憶する要素記憶部と、前記記憶された属性の列と、当該属性の列の間に埋め込む文字列の情報とを含むテンプレートを記憶するテンプレート記憶部と、前記入力を受け付けられた形態素列から、前記記憶された要素を検索し、当該検索結果の要素を列として出力する要素検索部と、前記検索結果の要素の列から当該要素の属性の列を取得し、前記記憶されたテンプレートのうち、当該要素の属性の列を含むものを用いて、当該検索された属性の列のそれぞれに対応する要素に対応する形態素列の間に、当該テンプレートが含む文字列の情報を埋め込んで、要約を生成する要約生成部とを備えることを特徴とするものである。

【特許文献1】特開2002-288190号公報

【発明の開示】

【発明が解決しようとする課題】

【0004】

前記背景技術の情報抽出を利用した文書要約装置によれば、ユーザが望む情報について文書を要約するのに好適な要約装置等を提供することができる。

しかしながら、テンプレートを利用した情報抽出を行う場合に、並列構造や入れ子構造など複雑な構成の文章が多数用いられていると、抽出項目が複数の文章にまたがって記述される場合があるなど、一般性が低くなる傾向がみられる。その結果、テンプレート数が増加し、多数の解候補があり抽出精度が低下する。さらに、大量テンプレートとのマッチングの処理によって、高速性が損なわれるなどの問題点がある。

【0005】

また、文書中には主要な製品情報以外に、製品の特徴を表す情報が出現している場合が多くある。また製品の特徴情報は記事中に複数出現している場合が多く、複数の特徴情報の中から特に重要度の高い特徴情報を抽出することが望まれる。

そこで、前記の課題を解決するために、指定された項目およびそれに関連する1つないしは複数個の情報、さらに指定された項目ではないが重要な情報を文書から見つけ出す高機能かつ高精度な情報抽出装置の提供を目的とする。

【課題を解決するための手段】

【0006】

まず、本発明における「売り」情報について定義する。次に、「売り」情報の抽出における記事見出しの有用性と、見出しと本文との関連性について述べる。

本発明における「売り」情報の定義について示すと、「売り」情報とは製品の特徴を示す情報の中で、特に重要度の高い情報である。しかし、記事中には複数の特徴情報が出現している場合が多く、各特徴情報に対する重要度は、特徴情報に対して重み付けを行う者の立場や観点によって大きく異なる。以下に、本発明で扱う新製品紹介記事と、記事の本文中に含まれる特徴情報の例を示す。

【0007】

図1は、本発明の実施形態に係る情報抽出装置における新製品紹介記事と記事中に含まれる特徴情報の説明図である。

見出し情報は、「[ビジネス情報] ニューロ制御のヒーター - - 三菱電機」である。本文情報は、「三菱電機は、設定温度を自動的に決めるニューロ制御の石油ファンヒーター

10

20

30

40

50

8タイプ24機種を8月21日発売する。6 16畳向けで価格は3万9000 8万9800円」である。ここで特徴情報は、「[設定温度を自動的に決める, ニューロ制御, 6 16畳向け]」となる。

これらの特徴情報は、重み付けの観点によってはどれも「売り」情報となり得る。従って、複数の特徴情報から「売り」情報を抽出するためには、この立場や観点を定める必要がある。そこで、本発明ではこの立場を新製品紹介記事の書き手、即ち記者と定め、記者の観点における「売り」情報を抽出する。

【0008】

[記事中の特徴情報について]

(a) 見出し中の特徴情報

記事の見出しとは、記者が読者に対し、本文の内容が一目で分かるようつけたものである。故に、見出し中に特徴情報が出現するとき、その特徴情報は記者の観点における重要度が特に高い「売り」情報であると判断できる。過去の分析用データ300記事のうち、見出しの特徴情報が製品の「売り」情報となっている記事は253記事であった。従って、「売り」情報を抽出する際に見出しを利用する事は極めて有効であるといえる。ここで、見出しの特徴情報が「売り」情報となっている記事の例を以下に示す。

【0009】

図2は、本発明の実施形態に係る情報抽出装置における見出し中に「売り」情報を含む記事の説明図である。

見出し情報は、「[雑記帳] 足利銀行が視覚障害者が利用できるATM」である。本文情報は、「足利銀行が沖電気工業と共同で全国で初めて視覚障害者が利用できるATMを開発した。従来のATMに、電話の受話器のようなハンドセットと取引金額や残高がボードに点字で浮き上がる点字表示装置、点字付き操作ボタンが加わった。操作すると、入金額や残高が点字で表示装置にでる。同行は「社会貢献活動の一環。より良い企業市民を目指します」とアピール。」である。ここで、「売り」情報は「視覚障害者が利用できる」となる。

【0010】

(b) 見出し中の特徴情報と本文との関連性

見出し中の特徴情報が本文中に出現するとき、本文中には見出し中の特徴情報の補足説明となる語句が含まれている記事が多くあった。従って、見出しと本文の両方に含まれる特徴情報に加え、その特徴情報を補足説明する語句を本文中より抽出することで、より詳細な「売り」情報の抽出が可能になる。本文中に見出しの特徴情報の補足説明となる情報を含む記事の例を以下に示す。

【0011】

図3は、本発明の実施形態に係る情報抽出装置における本文中に見出しの特徴情報の補足情報を含む記事の説明図である。

見出し情報は、「[ビジネス情報] 衛星放送チューナーを2台内蔵したテレビを発売 - シャープ」である。本文情報は、「シャープは、業界で初めて衛星放送チューナーを2台内蔵した29型カラーテレビ「ツインBS」を4月15日発売。22万円。衛星放送を見ながら、手持ちのVTRで衛星放送の裏番組録画が可能。同チューナーを内蔵していない別のテレビとケーブルで接続すれば、2つの衛星放送番組を同時に視聴できる。」である。ここで、補足情報は「業界で初めて」となる。

【0012】

(1) 本発明に係る情報抽出装置は、記事入力手段と、入力された記事情報についてテンプレートを利用して主要な製品情報を抽出するテンプレート抽出手段と、入力された記事情報について付与されたタグに基づいてパターンマッチングを行うダグパターンマッチング手段と、前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手段と、分割された記事の見出しを形態素解析する見出しの形態素解析手段と、形態素解析された見出しの文節から助詞を除去する見出し助詞除去手段と、前記テンプレートにより抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッ

10

20

30

40

50

チングする見出し特徴情報マッチング手段と、前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出手段と、分割された記事の本文を形態素解析する本文の形態素解析手段と、形態素解析された本文の文節から助詞を除去する本文助詞除去手段と、見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手段と、前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手段と、見出し特徴情報または本文特徴情報を売り情報として出力する売り情報の出力手段とを含むものである。

【 0 0 1 3 】

これにより、記事を入力し、入力された記事情報についてテンプレートを利用して主要な製品情報を抽出し、入力された記事情報について見出しと本文に分割し、見出しを形態素解析し、見出しの文節から助詞を除去し、前記テンプレートにより抽出された主要な製品情報と見出しの文節とをマッチングし、見出しの特徴情報を抽出し、記事の本文を形態素解析し、見出し特徴情報と本文の文節とをマッチングし、前記マッチングされた本文の特徴情報を抽出し、見出し特徴情報または本文特徴情報を売り情報として出力するので、定型性の高い文章に対しては簡易かつ迅速に抽出することができる。また、文書の中からあらかじめ定められた種類の情報やあらかじめ定められた種類の情報に関連する重要な情報が簡潔な言語表現で迅速かつ簡易に抽出可能となる。

【 0 0 1 4 】

(2) 本発明に係る情報抽出装置は、記事入力手段と、入力された記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出手段と、入力された記事情報について付与されたタグに基づいてパターンマッチングを行うダグパターンマッチング手段と、前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手段と、分割された記事の見出しを形態素解析する見出しの形態素解析手段と、形態素解析された見出しの文節から助詞を除去する見出し助詞除去手段と、前記係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチング手段と、前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出手段と、分割された記事の本文を形態素解析する本文の形態素解析手段と、形態素解析された本文の文節から助詞を除去する本文助詞除去手段と、見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手段と、前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手段と、見出し特徴情報または本文特徴情報を売り情報として出力する売り情報の出力手段とを含む。

【 0 0 1 5 】

これにより、記事を入力し、入力された記事情報について係り受け解析を利用して主要な製品情報を抽出し、入力された記事情報について見出しと本文に分割し、見出しを形態素解析し、前記係り受け解析により抽出された主要な製品情報と見出しの文節とをマッチングし、見出しの特徴情報を抽出し、記事の本文を形態素解析し、見出し特徴情報と本文の文節とをマッチングし、前記マッチングされた本文の特徴情報を抽出し、見出し特徴情報または本文特徴情報を売り情報として出力するので、定型性の低い複雑な文章に対しても簡易かつ精度よく抽出ができる。また、文書の中からあらかじめ定められた種類の情報やあらかじめ定められた種類の情報に関連する重要な情報が簡潔な言語表現で迅速かつ簡易に抽出可能となる。

【 0 0 1 6 】

(3) 本発明に係る情報抽出装置は、記事入力手段と、抽出精度の重み付けの閾値が一定以上のテンプレートを利用して入力された記事情報から主要な製品情報を抽出するテンプレート抽出手段と、前記テンプレートにより抽出されなかった記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出手段と、入力された記事情報について付与されたタグに基づいてパターンマッチングを行うダグパターンマッチング手段と、前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手段と、分割された記事の見出しを形態素解析する見出しの形態素解析手段と、形態素解析され

10

20

30

40

50

た見出しの文節から助詞を除去する見出し助詞除去手段と、前記テンプレートまたは係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチング手段と、前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出手段と、分割された記事の本文を形態素解析する本文の形態素解析手段と、形態素解析された本文の文節から助詞を除去する本文助詞除去手段と、見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手段と、前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手段と、見出し特徴情報または本文特徴情報を売り情報として出力する売り情報の出力手段とを含む。

【0017】

これにより、記事を入力し、抽出精度の重み付けの閾値が一定以上のテンプレートを利用して入力された記事情報から主要な製品情報を抽出し、テンプレートにより抽出されなかった記事情報について係り受け解析を利用して主要な製品情報を抽出し、記事を見出しと本文に分割し、記事の見出しを形態素解析し、前記テンプレートまたは係り受け解析により抽出された主要な製品情報と見出しの文節とをマッチングし、見出しの特徴情報を抽出し、本文を形態素解析し、見出し特徴情報と本文の文節とをマッチングし、前記マッチングされた本文の特徴情報を抽出し、見出し特徴情報または本文特徴情報を売り情報として出力するので、テンプレート抽出により定型性の高い文に対してのみ抽出でき、テンプレートマッチでの誤った抽出の減少、および処理時間の短縮ができる。また、予めテンプレートによる抽出を行うことで、係り受け解析の負担を軽減し、全体の抽出精度の向上を図ることができる。そのうえ、文書の中からあらかじめ定められた種類の情報やあらかじめ定められた種類の情報に関連する重要な情報が簡潔な言語表現で高機能性及び高性能性を保ちながら、迅速かつ簡易に抽出可能となる。

【0018】

(4) 本発明に係る情報抽出装置は必要に応じて、前記形態素解析された本文の係り受け関係を調べる本文の係り受け関係解析手段と、係り受け関係により修飾語句を補足説明情報として抽出する補足説明の抽出手段と、補足説明情報を売り情報として出力する売り情報の出力手段とを含む。

これにより、形態素解析された本文の係り受け関係を調べ、係り受け関係により修飾語句を補足説明情報として抽出し、補足説明情報を売り情報として出力するので、あらかじめ定められた種類の情報やあらかじめ定められた種類の情報に関連する重要な情報だけでなく、あらかじめ定められた種類の情報ではないが、興味深い、重要だと思われる情報が簡潔な言語表現で迅速かつ簡易に抽出可能となる。また、少ない情報量でより詳しい製品の情報を得ることができる。

【0019】

(5) 本発明に係る情報抽出装置は必要に応じて、前記テンプレート抽出手段は、入力された記事を句点ごとに分割する記事句点分割手段と、記事の1行目に対応するAテンプレート集合とマッチングを行うAテンプレートマッチング手段と、前記Aテンプレート集合によりマッチングされた製品の特徴情報を抽出するAテンプレート抽出手段と抽出された製品の特徴情報について抽出項目ごとにチェックする制約チェック手段と、情報を抽出することができたテンプレートのIDを記憶するテンプレートID記憶手段と、記事の2行目以降に対応するBテンプレート集合とマッチングを行うBテンプレートマッチング手段と、前記Bテンプレート集合によりマッチングされた製品の特徴情報を抽出するBテンプレート抽出手段と、抽出された製品の特徴情報である抽出解と製品を対応付けるテンプレート製品対応手段とを含む。

【0020】

これにより、入力された記事を句点ごとに分割し、記事の1行目に対応するAテンプレート集合とマッチングを行い、前記Aテンプレート集合によりマッチングされた製品の特徴情報を抽出し、抽出された製品の特徴情報について抽出項目ごとに制約をチェックし、情報を抽出することができたテンプレートのIDを記憶し、記事の2行目以降に対応する

10

20

30

40

50

Bテンプレート集合とマッチングを行い、前記Bテンプレート集合によりマッチングされた製品の特徴情報を抽出し、抽出された製品の特徴情報である抽出解と製品を対応付けるので、分野が限定され、かつ、文章構造が単純な文書からの情報抽出処理に対しては、全文の構文要素を解析せず、表層の単語列の並びに現れる特定のパターンを認識することから簡易かつ迅速に抽出を行うことができる。

【0021】

(6)本発明に係る情報抽出装置は必要に応じて、前記係り受け抽出手段は、入力された記事に付与されたタグに基づいてパターンマッチングを行う係り受けタグパターンマッチング手段と、前記パターンマッチングの結果に基づいて記事を見出しと本文に分割する係り受けタグ分割手段と、分割された見出しに含まれる特殊記号を分析する見出し分析手段と、前記見出しに含まれる特殊記号の後方の語句を「販売元」情報として処理する見出し処理手段と、分割された本文を句点ごとに分割する本文句点分割手段と、前記本文中に括弧内数値が存在するか否かを判定する括弧内数値判定手段と、括弧内数値が存在すると判断した場合に、構文解析により文節情報を作成する文節情報作成手段と、括弧内数値が存在しないと判断した場合に、固定パターンが存在するか否かを判断する固定パターン判定手段と、固定パターンが存在すると判断した場合に、固定パターンと文節情報から得られる固定パターンの係り受け情報を利用して固定パターンに係る文節情報集合を作成する固定パターン係り受け作成手段と、前記作成された固定パターンに係る文節情報集合から固定パターン及び各形式について定めた条件に従って文節情報を抽出する抽出手段と、抽出された文節情報から不要な情報を削除して抽出解作成手段と、抽出解から製品に対する対応や割り当てを行う係り受け対応・割付手段とを含む。

【0022】

これにより、入力された記事に付与されたタグに基づいてパターンマッチングを行い、前記パターンマッチングの結果に基づいて記事を見出しと本文に分割し、分割された見出しに含まれる特殊記号を分析し、前記見出しに含まれる特殊記号の後方の語句を「販売元」情報として処理できる。また、分割された本文を句点ごとに分割し、前記本文中に括弧内数値が存在するか否かを判定し、括弧内数値が存在すると判断した場合に、構文解析により文節情報を作成し、括弧内数値が存在しないと判断した場合に、固定パターンが存在するか否かを判断し、固定パターンが存在すると判断した場合に、固定パターンと文節情報から得られる固定パターンの係り受け情報を利用して固定パターンに係る文節情報集合を作成することができる。そして、前記作成された固定パターンに係る文節情報集合から固定パターン及び各形式について定めた条件に従って文節情報を抽出し、抽出された文節情報から不要な情報を削除して抽出解を作成し、抽出解から製品に対する対応や割り当てを行うことができるので、抽出情報はある特定の文節に係るという点に着目して抽出を実現することができる。文章構造が複雑で、文書表層の単語列の並びに特定のパターンがなくとも抽出できる。

【0023】

(7)本発明に係る情報抽出方法は、記事入力ステップと、入力された記事情報についてテンプレートを利用して主要な製品情報を抽出するテンプレート抽出ステップと、前記テンプレートにより抽出されなかった記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出ステップと、入力された記事情報について付与されたタグに基づいてパターンマッチングを行うタグパターンマッチングステップと、前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割ステップと、分割された記事の見出しを形態素解析する見出しの形態素解析ステップと、形態素解析された見出しの文節から助詞を除去する見出し助詞除去ステップと、前記テンプレートまたは係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチングステップと、前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出ステップと、分割された記事の本文を形態素解析する本文の形態素解析ステップと、形態素解析された本文の文節から助詞を除去する本文助詞除去ステップと、見出し特徴情報と形態素解析後に助詞

10

20

30

40

50

を除去された本文の文節とをマッチングする本文特徴情報マッチングステップと、前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出ステップと、形態素解析された本文の係り受け関係を調べる本文の係り受け関係解析ステップと、係り受け関係により修飾語句を補足説明情報として抽出する補足説明の抽出ステップと、見出し特徴情報または本文特徴情報または補足説明情報を売り情報として出力する売り情報の出力ステップとを含む。

【 0 0 2 4 】

(8) 本発明に係る情報抽出プログラムは、記事入力手順と、入力された記事情報についてテンプレートを利用して主要な製品情報を抽出するテンプレート抽出手順と、前記テンプレートにより抽出されなかった記事情報について係り受け解析を利用して主要な製品情報を抽出する係り受け抽出手順と、入力された記事情報について付与されたタグに基づいてパターンマッチングを行うダグパターンマッチング手順と、前記パターンマッチング結果に基づいて記事を見出しと本文に分割する記事分割手順と、分割された記事の見出しを形態素解析する見出しの形態素解析手順と、形態素解析された見出しの文節から助詞を除去する見出し助詞除去手順と、前記テンプレートまたは係り受け解析により抽出された主要な製品情報と形態素解析後に助詞を除去された見出しの文節とをマッチングする見出し特徴情報マッチング手順と、前記主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する見出し特徴情報抽出手順と、分割された記事の本文を形態素解析する本文の形態素解析手順と、形態素解析された本文の文節から助詞を除去する本文助詞除去手順と、見出し特徴情報と形態素解析後に助詞を除去された本文の文節とをマッチングする本文特徴情報マッチング手順と、前記マッチングされた本文の特徴情報を抽出する本文特徴情報抽出手順と、形態素解析された本文の係り受け関係を調べる本文の係り受け関係解析手順と、係り受け関係により修飾語句を補足説明情報として抽出する補足説明の抽出手順と、見出し特徴情報または本文特徴情報または補足説明情報を売り情報として出力する売り情報の出力手順としてコンピュータを機能させる。

これら前記の発明の概要は、本発明に必須となる特徴を列挙したのではなく、これら複数の特徴のサブコンビネーションも発明となり得る。

【 発明を実施するための最良の形態 】

【 0 0 2 5 】

ここで、本発明は多くの異なる形態で実施可能である。したがって、下記の実施形態の記載内容のみで解釈すべきではない。実施形態では、主に装置について説明するが、所謂当業者であれば明らかな通り、本発明は、コンピュータで使用可能なプログラムとしても実施できる。また、本発明では、ハードウェア、ソフトウェア、または、ソフトウェア及びハードウェアの実施形態で実施可能である。プログラムは、ハードディスク、CD ROM、DVD-ROM、光記憶装置または磁気記憶装置等の任意のコンピュータ可読媒体に記録できる。さらに、プログラムはネットワークを介した他のコンピュータに記録することが出来る。

【 0 0 2 6 】

(本発明の第 1 の実施形態)

[1 . ハードウェア構成]

図 4 に本発明の実施形態における情報抽出装置のハードウェア構成図を示す。コンピュータ 1 は、例えば、CPU (Central processing Unit) 2、メインメモリ 3、HDD (Hard Disk Drive) 4、ビデオカード 5、マウス 6、キーボード 7、光学ディスク 8 等により構成される。なお、本実施形態においては、図 4 に示すように、情報抽出装置を一のコンピュータ上に構築した例を説示するが、クライアントであるウェブブラウザ装置とサーバである WWW サーバからなる WWW システムを利用して情報抽出機能をブラウザ装置上で使用する構成とすることは所謂当業者であれば明らかである。例えば、WWW サーバに情報抽出機能をアドインとして実装する。または、ウェブブラウザに情報抽出機能をプラグインとして実装することもできる。さらに、

WWWサーバに情報抽出機能の一部をアドインとして、ウェブブラウザに残りの情報抽出機能をプラグインとして実装することもできる。

【 0 0 2 7 】

[2 . ブロック構成]

図5は、本発明の実施形態に係る情報抽出装置のブロック構成図である。情報抽出装置は、入力部10、テンプレート抽出部20、係り受け抽出部30、タグパターンマッチング部40、記事分割部50、見出しの形態素解析部60、見出しの助詞除去部70、見出し特徴情報マッチング部80、見出し特徴情報抽出部90、本文の形態素解析部100、本文の助詞除去部110、本文特徴情報マッチング部120、本文特徴情報抽出部130、本文の係り受け解析部140、補足説明の抽出部150、売り情報の出力160を含む。タグパターンマッチング部40は、記事にタグを付け、タグのパターンマッチングを行う。記事分割部50は、記事をタグのパターンマッチの結果に従い、見出しと本文に分割する。見出しの形態素解析部60は、分割された見出しの形態素解析を行う。ここで、形態素解析は、形態素解析システムJUMANを利用することができる。JUMANとは、日本語の形態素解析を行うためのシステムで、日本語の文章を入力とし、入力文を単語単位に区切り、それぞれの形態素を決定するものである。図6は、本発明の実施形態に係る情報抽出装置における形態素解析結果である。入力文は「セガ・エンタープライゼス社は、SF映画的な光線銃戦を模擬体験できるおもちゃ「ロックオン」を発売した」である。見出しの助詞除去部70は、形態素解析したのちに見出しに含まれる助詞を除去する。見出し特徴情報マッチング部80は、テンプレート及び係り受け解析により抽出された製品情報と見出しに含まれる特徴情報のマッチングを行う。本文の形態素解析部100及び本文の助詞除去部110は、見出しの形態素解析及び助詞除去と同様である。本文特徴情報マッチング部120は、形態素解析後に助詞を除去した本文と見出しの特徴情報のマッチングを行う。本文特徴情報抽出部130は、マッチングした本文の中から特徴情報を抽出する。本文の係り受け解析部140は、係り受け解析器を用いて本文の係り受けを調べる。ここで、構文解析には構文解析システムKNPを用いることができる。KNPとは、日本語の構文解析を行うためのシステムで、JUMAN出力結果を入力とし、それらを文節単位にまとめ、文節間の係り受け関係を決定するものである。図7は、本発明の実施形態に係る情報抽出装置における係り受け解析結果である。入力文は、「セガ・エンタープライゼス社は、SF映画的な光線銃戦を模擬体験できるおもちゃ「ロックオン」を発売した」である。補足説明の抽出部150は、係り受け解析の結果、修飾語句を補足説明情報として抽出する。売り情報の出力160は、抽出された特徴情報を出力する。

【 0 0 2 8 】

[3 . 動作]

図8は、本発明の実施形態に係る情報抽出装置における処理フローシートである。まず、入力部10が記事を入力する(S100)。ここで、テンプレート抽出部20がテンプレートによる抽出処理をする(S200)。なお、テンプレートによる抽出処理については、後述する。抽出できたか否か判断する(S300)。ここで、テンプレートによる抽出を、定型性の高い文に対してのみ抽出できるようにするために、テンプレートに対し抽出精度に基づいた重み付けを行い、この重みがある閾値を超えるもののみを利用する。この重みを利用することで、テンプレートによる抽出は、抽出精度の高い文に対してのみ行うことができる。これは、テンプレートマッチでの誤った抽出の減少、および処理時間の短縮に繋がる。また、予めテンプレートによる抽出を行うことで、係り受け解析の負担を軽減し、全体の抽出精度の向上を図ることができる。抽出が出来ていなければ、係り受け抽出部30が係り受け解析による抽出を行う(S400)。係り受け抽出処理についても後述する。また、入力された記事については、予め索引番号、見出し、本文にあたる部分にそれぞれ対応するタグがつけられているのでタグパターンマッチング部40がタグのパターンマッチングを行う(S500)。そして、記事分割部50が記事を見出しと本文に分割する(S600)。見出しの形態素解析部60が、見出しの形態素解析を行い文節に区切る(S700)。見出しの助詞除去部70が各文節から助詞を取り除く(S800)

。見出し特徴情報マッチング部 80 がテンプレート及び係り受けを利用して抽出された主要な製品情報と見出しの文節とのマッチングを行う (S 900)。見出し特徴情報抽出部 90 が主要な製品情報以外の情報を示す文節を見出しの特徴情報として抽出する (S 1000)。分割された本文は、本文の形態素解析部 100 が本文の形態素解析を行い文節に区切る (S 1100) 本文の助詞除去部 110 が助詞の除去をする (S 1200)。本文特徴情報マッチング部 120 が各文節と見出しの特徴情報とのマッチングを行う (S 1300)。本文特徴情報抽出部 130 が本文から見出しの特徴情報を示す文節を抽出する (S 1400)。なお、見出しの特徴情報の同義語にあたる語句を抽出するため、同義語辞書を参照することができる。入力された記事の本文について本文の係り受け解析部 140 が本文の構文解析を行い、本文より抽出した文節の係り受け関係を調べる (S 1500)。補足説明の抽出部 150 が抽出した文節を修飾する語句があれば、それを特徴情報の補足情報として抽出する (S 1600)。売り情報の出力 160 が売り情報の出力を行う (S 1700)。ここで、売り情報は、まとめて出力することも、見出しの特徴情報、本文の特徴情報、補足説明情報と必要に応じて個別に出力することもできる。

【0029】

[4. テンプレートによる抽出について]

ここで、テンプレート抽出処理について、詳細を説明する。

図9は、本発明の実施形態に係る情報抽出装置におけるテンプレート抽出のブロック構成図である。テンプレート抽出部 20 は、記事句点分割部 210、Aテンプレートマッチング部 220、Aテンプレート抽出部 230、制約チェック部 240、テンプレート ID 記憶部 250、Bテンプレートマッチング部 260、Bテンプレート抽出部 270、テンプレート対応・割付部 280 を含む。記事句点分割部 210 は、入力された記事を、句点ごとに分割する。Aテンプレートマッチング部 220 は、記事の 1 行目に対応するテンプレートによるマッチングを行う。Aテンプレート抽出部 230 は、1 行目に対応するテンプレートによるマッチングの抽出を行う。制約チェック部 240 は、抽出項目ごとに制約チェックを行う。なお、抽出項目に対する制約には抽出項目に依存する制約と抽出項目に依存しない制約がある。抽出項目に依存する制約とは、例えば「製品種別」、「販売元」、「価格」、「販売日」などを含む。具体的には、「製品種別」に関するテンプレートからの抽出文に対して、例えば、「丸括弧に含まれる文字列の除去」、単語の区切りが間違っている解候補の除去、「解候補の品詞並びから品詞が「名刺」、「接頭辞」など以外の品詞が含まれていたら除去」などである。抽出項目に依存しない制約とは、抽出項目の性質とは関係なく、明らかに意味のない句を排除する。例えば、「括弧の対応がついているか」、「読点から始まっているか」、「例えば「や」や「あ」などの禁則開始文字で始まっているか」などである。テンプレート ID 記憶部 250 は、抽出されたテンプレートの ID を記憶する。Bテンプレートマッチング部 260 は 2 行目以降に対応するテンプレートによるマッチングを行う。Bテンプレート抽出部 270 は、2 行目以降に対応するプレートによるマッチングの抽出を行う。テンプレート製品対応部 280 は、抽出解における対応を行う。抽出解における対応は、抽出項目の出現パターンに応じて、予め設定された抽出項目間の関係に基づいて行われる。例えば、「製品名」または「製品の細分類」が複数出現し、「価格」、「発売日」が単数で出現する場合に複数の項目が単数の項目に対応する。具体的には、ビール、ジュース等の複数の製品が 110 円で 8 月 13 日に販売されるなどである。また、「価格」、「販売日」が複数で出現する場合は、複数の項目同士が対応する場合などもある。さらに、項目間において「製品名」が単数で「製品の細分類」が複数である場合では、「製品名」がすべての「製品の細分類」に対応するなど、項目間に上位と下位の関係があることから、その間に対応関係があると予め設定して、対応付けを行う。

【0030】

テンプレートの定義について説明する。製品情報抽出に用いるテンプレートは、実験対象である新製品紹介記事の抽出すべき項目にタグを付与したデータから、(1) 抽出項目が出現する文章に頻出する表現、(2) 抽出項目前後の形態素、(3) 抽出項目の種類

10

20

30

40

50

3つの情報を残したものである。なお、抽出項目は、例えば、「販売元」、「製品種別」、「製品名」、「製品の細分類」、「価格」、「発売日」などを含む。ここで、テンプレート抽出処理において予め行われるテンプレートの作成及びテンプレートの重み付けの前処理について以下に説明する。

【0031】

図10は、本発明の実施形態に係る情報抽出装置における分析用タグ付きデータ例の説明図である。この例のように複数の製品を紹介する記事には、各製品毎に抽出項目の対応を“ I = < . . . > - < . . . > . . . ” のリストの形で付与する。具体的には、図10に示すように、「 < c 1 > 富士通ゼネラル < / c 1 > は横長でワイド感のある画面を用いた < k 1 > 29型衛星放送 (BS) 内臓テレビ < / k 1 > 「 < n 1 > BS - 29M55 < / n 1 > 」と「 < n 2 > BS - 29M50 < / n 2 > 」を < d 1 > 9月2日 < / d 1 > に発売する」に対して、 I = < c 1 , d 1 , k 1 , n 1 , 0 , 0 > - < c 1 , d 1 , k 1 , n 2 , 0 , 0 > となる。

10

【0032】

そして、タグ付きデータを用いて抽出項目が存在する文章に頻出する表現をまとめ、「固定パターン」とする。この固定パターンとタグの前後1形態素を残し、それ以外をワイルドカードとして、任意の文字列がマッチできるようにして、テンプレートを作成する。なお、形態素の切り出しには、形態素解析器 JUMAN を用いることができる。上記のタグ付きデータからテンプレートを作成する手順を次に示す。まず、タグ付きデータの「 < tag > . . . < / tag > 」箇所を、それぞれのタグが表す抽出項目に置き換える。例えば、{販売元1}は横長でワイド感のある画面を用いた{製品種別1}「{製品名1}」と「{製品名2}」を{発売日1}に発売する。次に、抽出項目の前後1形態素と固定パターン(発売する。)以外をワイルドカードに置き換える。例えば、{販売元1}は*用いた{製品種別1}「{製品名1}」と「{製品名2}」を{発売日1}に発売する。完成されたテンプレートについては、この例のように、1文中1つの項目に複数の情報が存在する場合は、次の対応リストを付与する。1つの製品の場合は不要である。

20

- {販売元1} - {製品種別1} - {製品名1} - {発売日1}
- {販売元1} - {製品種別1} - {製品名2} - {発売日1}

【0033】

続いて、テンプレートの優先順位付けについて説明する。実際の抽出処理では、テンプレート集合と記事とのマッチングを行うため、一意に解が決まることは殆んどなく、複数の解候補が存在する。この解候補に対して優先順位付を行い最も優先順位が高いものをその記事の抽出情報とする。記事全てのマッチング処理が終了した後、その解候補全てについて優先順位付けを行うこともできるし、一文毎に優先順位付けを行うこともできる。ここでは、1文毎の場合について説明する。優先順位付けは予めテンプレートに重みを与えておき、その情報を利用して行う。重みは以下の簡易な方法により付与するので以下にその説明をする。(1)テンプレート集合の各テンプレートを今回作成した情報抽出システムを用いてテンプレート作成用データにマッチングさせ、マッチした文の数およびマッチしてかつ情報抽出が成功した文の数を記憶しておく。(2)以下の式で重みを決定する。

30

【0034】

【数1】

$$\text{重み} = \frac{\text{マッチしてかつ情報抽出が成功した文の数}}{\text{マッチした文の数}} \times \text{抽出する抽出情報の個数}$$

40

重み付けの例を、次の記事例について述べる。

【0035】

【表 1】

<s1>車椅子 2 台用</s1>と<s2> 1 台用</s2>の 2 タイプを用意、
<s1> 1 台用</s1>で価格は<p1> 2 5 0 万円</p1> (消費税含まず)。

テンプレートは、それぞれ<s 1 >と<s 2 >の 2 種類で、<s 1 >は*、<p 1 > は p 1 であるとし、頻度は、マッチした回数 2 0 回、正解を返した回数 1 0 回とすると、抽出する抽出情報は s 1 , s 2 , s 1 , p 1 の 4 つであるので、以下の式が与えられる。

【 0 0 3 6 】

10

【数 2】

$$\text{重み} = \frac{10}{20} \times 4 = 2$$

【 0 0 3 7 】

一般には s 1 , s 2 , p 1 , p 2 あるいは s 1 , p 1 , s 2 , p 2 の形をとるものが多いが、形はさまざまである。例えば、s 1 がさらに細分化されており、s 1 (s 3 , p 3 , s 4 , p 4) , s 2 , p 2 のように s 3 , p 3 , s 4 , p 4 が埋め込まれていて s 1 に対応する p 1 がもともとない場合もある。このような場合でも s 3 , p 3 , s 4 , p 4

20

から計算される値を使うなど、上式を拡充して用いることができる。このようにして、定義したテンプレートを、学習データより大量に作成し、実験データの入力とマッチングさせて抽出を行う。なお、テンプレート集合は予め「テンプレート作成データの 1 文目から作成されたテンプレート集合」(A プレートとする) と「2 文目以降から作成されたテンプレート集合」(B プレートとする) に分類する。

【 0 0 3 8 】

図 1 1 は、本発明の実施形態に係る情報抽出装置におけるテンプレート抽出処理のフローシートである。まず、記事句点分割部 2 1 0 が、入力記事を区点で分割する (S 2 0 1) 。 A テンプレートマッチング部 2 1 0 本文の 1 文目に対して「1 文目から作成されたテンプレート集合」(A プレート) とマッチングを行う (S 2 0 2) 。 A テンプレート抽出部 2 3 0 がマッチングに対する抽出を行う (S 2 0 3) 。テンプレートによって抽出された文字列を抽出項目ごとに制約チェック部 2 4 0 が制約チェックを行う (S 2 0 4) 。制約チェックを全てクリアした場合、テンプレート ID 記憶部 2 5 0 がその抽出情報を抽出したテンプレートの属するテンプレート集合の ID を記憶する (S 2 0 5) 。次の文があるか否かを判断する (S 2 0 6) 。次の文があると判断した場合に、B テンプレートマッチング部 2 6 0 が先程記憶したテンプレート ID の従属する「2 文目以降から作成されたテンプレート集合」のテンプレート ID と次の文とのマッチングを行う (S 2 0 7) 。B テンプレート抽出部 2 7 0 がマッチングに対する抽出を行う (S 2 0 8) 。そして、制約チェック部 2 4 0 がテンプレートによって抽出された文字列を抽出項目ごとに制約チェックを行うに戻る (S 2 0 4) 。次の文がないと判断した場合に、テンプレート製品対応部 2 8 0 がテンプレートに付与された対応リストに従って抽出解を製品に対応付ける (S 2 0 9) 。

30

40

【 0 0 3 9 】

[5 . 係り受け解析による抽出について]

係り受け解析による抽出について以下に説明する。図 1 2 は、本発明の実施形態に係る情報抽出装置における係り受け解析による抽出のブロック構成図である。係り受け解析抽出部 3 0 は、係り受けタグパターンマッチング部 3 1 0、係り受けタグ分割部 3 2 0、見出し分析部 3 3 0、見出し処理部 3 4 0、本文句点分割部 3 5 0、括弧内数値判定部 3 6 0、文節情報作成部 3 7 0、固定パターン判定部 3 8 0、固定パターン係り受け作成部 3 9 0、係り受け抽出部 4 0 0、抽出解作成部 4 1 0、係り受け対応・割付部 4 2 0 を含む

50

。なお、構文解析には前述した日本語構文解析システムKNPを用いることができる。係り受けタグパターンマッチング部310は、記事にタグを付け、タグのパターンマッチングを行う。係り受けタグ分割部320は、タグのパターンマッチの結果に従い、見出しと本文、記事終了に判定、分割する。見出し分析部330は、見出しに含まれる特殊記号を分析する。見出し処理部340は、特殊記号の後ろにある語句を「販売元」として利用する。本文句点分割部350は、分割された本文を、句点で分割する。括弧内数値判定部360は、括弧内の数値があるか否かを判定する。文節情報作成部370は、括弧内数値がある場合に、構文解析により文節情報を作成する。ここで、構文解析には前述した構文解析システムKNPを用いることができる。固定パターン判定部380は、括弧内数値がない場合に、固定パターンがあるか否かを判定する。固定パターン係り受け作成部390は、文節情報と固定パターンから固定パターン係り受けを作成する。係り受け抽出部400は、係り受け関係を抽出する。抽出解作成部410は、重複要素等の削除を行い、抽出解を作成する。係り受け対応・割付部420は、抽出解における製品との対応や割付を行う。

10

ここで、抽出情報の係り受けを調べるために、学習データの分析より、抽出情報を受ける文節、「固定パターン」とその固定パターンに係る抽出情報の「格形式」を定義する。

【0040】

図13は、本発明の実施形態に係る情報抽出装置における固定パターンと格形式の説明図である。1文目について以下に述べる。固定パターンが、「発売」、「販売」、「売り」、「チェンジ」、「開発」、「改良」、「開始」、「始め」、「発表」、「商品化」、「製品化」、「輸入」、「発刊」、「発行」、「出版」、「創刊」、「刊行」については、抽出情報とみなす格形式が、“未格，ガ格”：c、“カラ格，隣接，無格，二格”：d、“ヲ格”：knsとする。固定パターンが、「展開」、「参入」、「提携」については、抽出情報とみなす格形式が、“未格，ガ格：c”、“カラ格，隣接，無格，二格”：dとする。

20

固定パターンが、「変更」、「強化」、「強調」については、抽出情報とみなす格形式が、“ヲ格<-ノ格”：knsとする。固定パターンが、「採用」、「導入」については、“ヲ格<-二格”：knsとする。固定パターンが、「追加」、「設定」、「加え」について、“二格，ヲ格”：knsとする。固定パターンが、「搭載」、「装備」について、“体言でノ格，用言で二格”：knsとする。

30

【0041】

2文目以降について以下に述べる。固定パターンが、「[製品数]（文末）」については、抽出情報とみなす格形式が、“ノ格，同格未格”：knsとする。固定パターンが、「発売（文末）」については、抽出情報とみなす格形式が、“未格”：knsとする。固定パターンが、「発売（文頭）」については、抽出情報とみなす格形式が、受け文節：knsとする。固定パターンが、「別売り（ノ格）」については、抽出情報とみなす格形式が、受け文節：knsとする。固定パターンが、「別売り（ノ格以外）」については、“デ格”：knsとする。

【0042】

その他について以下に述べる。固定パターンが、「[価格表記]」については、抽出情報とみなす格形式が、“ガ格，未格，デ格，隣接”：knsとする。固定パターンが、「[「」]」については、抽出情報とみなす格形式が、“同格連体，ノ格”：knsとする。固定パターンが、「[製品数]，新製品，新商品，新モデル（等）」については、抽出情報とみなす格形式が、“ノ格，同格未格”：knsとする。

40

ここで、解候補の抽出ルールについて、以下に説明する。図13の固定パターンと格形式を用いて固定パターンに係る文節から解候補を選別し、個々の項目によって詳細ルールを定める。このルールにより係り受けで得た解候補集合に新たな解候補を追加したり、解候補集合から不必要な解候補の削除したりする作業を行い、抽出結果を作成する。

【0043】

(a) 「販売元」について

50

販売元を含む文節の格形式は主に「未格，ガ格」であり、以下の条件を満たすものを販売元の解候補とする。(1)第1文目、能動態述語文節に“未格”または“ガ格”で係る文節。(2)第1文目、受動態の述語文節に“カラ格”で係る文節。ただし、複数販売元が共同で製品を開発している場合は、特に、(3)(1)及び(2)で販売元解候補と名詞並列の文節。(4)“共同”を含む文節に“ト格”で係る文節という条件が追加される。

【0044】

(b)「発売日」について

発売日を含む文節の格形式は主に「カラ格，二格，無格，隣接」であり、以下の条件を満たすものを発売日の解候補とする。(1)製品の発売を表現する第1文目、能動態の述語文節に“カラ格，二格，無格，隣接”のいずれかで係る文節。(2)製品の発売を表現する第1文目、受動態の述語文節に“二格，無格，隣接”で係る文節。(3)(1)及び(2)から発売日解候補が得られない場合、それ以降の文で“発売”に(1)の格形式で係る。または、日付表記の文末。ただし、複数製品が異なる発売日に発売される場合は、特に、(4)(1)の条件を満たす文節が述語並列になっている場合、並列範囲内の日付表現の文節。(5)2文目以降の“発売”に(1)の格形式で係る、または、日付表記の文末という条件が追加される。(1)～(5)で用いている「日付表記」とは、“数値＋「月」数値＋「日」”や“「上旬」、「中旬」、「下旬」”などの日付表記パターンで、これらをまとめた正規表現を作成している。

【0045】

(c)「製品種別・製品名・細分類」について

「販売元」、「発売日」は固定パターンとの係り受けで抽出可能であるが、「製品種別・製品名・細分類」は、固定パターンから抽出できる解候補とその解候補周辺の係り受けで解候補を得る。解候補がそろった段階で項目に割り当てるため、ここでは3項目を一括して条件をまとめる。ここで、第1文目の固定パターンからの解候補は、(1)“発売”に類する固定パターンに“ヲ格”で係る文節。(2)“追加、設定”に類する固定パターンに“ヲ格，二格”で係る文節。(3)“採用、導入、強化、強調”に“二格”で係る文節。(4)“装備、搭載、刊行”に類する固定パターンに体言文節は“ノ格”，用言文節は“二格”で係る(1)～(4)で抽出した解候補について、(5)解候補が鉤括弧を含む場合、解候補に係る文節の格形式が“同格連体”または“ノ格”であれば解候補に追加する。(6)解候補が「機種」など製品数を表す場合は、その候補を削除し、製品数に係る文節を解候補に追加する。また、2文目以降の固定パターンからの解候補は、(7)製品数を表す文末に“ノ格，同格未格”で係る文節。(8)「別売」を表す文節が“ノ格”であった場合、これを受ける文節。(9)「別売」を表す文節が“デ格”であった場合、これに係る文節。(10)文頭が「発売」を表す文節の場合、これを受ける文節。(11)文末が「発売」を表す固定パターンであった場合、“未格(助詞：モ)”で係る文節。さらに、価格表記を含む文節からの解候補は、(12)括弧なしの価格表記の場合、“未格、ガ格、デ格、隣接”で係る文節。(13)括弧ありの価格表記の場合、“同格連体、連体”に係る文節。それぞれ解候補は名詞並列であった場合は、解候補と同等の文節も解候補とする。

【0046】

(d)「価格」について

「価格」は他の5項目の抽出法と異なり、表記パターンによるパターンマッチで抽出を行う。しかし、価格表記であっても、他の製品と比較した差額や、売上目標を表すものがある。そのため、パターンマッチを行った後それらのパターンをまとめて制約をかけ、解候補とする。なお、解候補とみなさない価格表記の制約を以下に示す。(1)価格表記直前に以下の表現を含むものは解候補としない。例えば、「売上高」、「売上目標」、「コスト」、「年間」、「電気代」、「ガス代」、「資金」などである。(2)価格表記直後に以下の表現を含むものは解候補としない。例えば、「割安」、「下回」、「切る」、「安く」、「低価格」、「下げる」、「値下げ」、「引き下げ」、「値引き」、「安い」、

10

20

30

40

50

「低い」、「抑える」、「とどめる」、「価格削減」、「割増」、「アップ」、「高い」、「値上げ」、「市場規模」、「売り上げ」、「費用」、「電気代」、「ガス代」、「資金」、「売上高」、「年間売上」などである。

【0047】

図14は、本発明の実施形態に係る情報抽出装置における係り受け抽出処理のフローシートである。まず、係り受けタグパターンマッチング部310が、タグのパターンマッチを行う(S401)。ここで、記事には(1)“<索引記事番号>”、(2)“<詳細画面用記事見出し>”、(3)“<記事全文>”、(4)“</記事全文>”の4つのタグが付与されており、文が入力される度に、このタグのパターンマッチングによって、係り受けタグ分割部320が、「見出し」と、「本文」、「記事終了」を判定できる(S402)。「見出し」の場合はS403へ、「本文」の場合はS405へ、「記事終了」の場合はS413へ渡す。「見出し」については、見出し分析部330が見出しの内容に特殊記号の“- - ”、“=”、“ ”のいずれかが含まれているか否かを分析する(S403)。見出しの内容に“- - ”、“=”、“ ”のいずれかが含まれている場合は、見出し処理部340がマッチした記号の後ろを販売元として利用する(S404)。「本文」については、本文句点分割部350が入力記事を区点で分割する(S405)。括弧内数値判定部360が括弧内価格の有無を判定する。ここで、分析用データから求めた価格表記を基に、以下の正規表現を作成した。

10

【0048】

【表2】

「((約|同|計)?([,、一二三四五六七八九十百千万億〇1234567890・]+円)+
(一|~|から|台|台から|ちょうど|安|以下|高|程度|程度から)?)+|未定|オープンプ
ライス|((同)?([,、一二三四五六七八九十百千万億〇1234567890・]+(円)?
+(一)([,、一二三四五六七八九十百千万億〇1234567890・]+円)+(程度|
台)?)」

20

【0049】

これにマッチするものを価格と判定する(S406)。括弧内価格にマッチしないと判定された場合に、固定パターン判定部380が固定パターンを分類別に配列に格納し、適宜本文とマッチングし、固定パターンを含む文であるかを判定する(S407)。固定パターンにマッチしない文は抽出処理を行わずS401へ戻る。括弧内価格にマッチすると判定した場合に、文節情報作成部370が文節情報を作成する(S408)。ここで、本文1文をKNPにかけることによって文節に分割することができる。この文節と、抽出・割り当て・対応付けに必要な、係り受け情報(各文節の係り先)、格情報(ガ格、ヲ格、同格連体など)、用言か体言か、名詞並列・述語並列の範囲、文中の文節の出現位置、情報を抜粋し、文節情報を作成することとなる。固定パターン係り受け作成部390が固定パターン係り受けを作成する(S409)。つまり、文節情報と固定パターンを受け取り、係り受け情報を利用して固定パターンに係る文節情報集合を作成する。

30

40

【0050】

図15は、本発明の実施形態に係る情報抽出装置における構文解析結果及びその文節情報の対応の説明図である。図15(a)のKNP構文木結果が図15(b)の文節情報に対応している。具体的には、図15(a)は、「 は、低コストの店舗監視記録用ビデオ「SR-L900」を2日に販売する」の構文である。図15(b)では、「2日に」は、「[3, “2日に”, [0, 0], [“二格”, “体言”], “5”]」となる。また「発売する。」は、「0, “発売する。”, [0, 0], [nil, “用言”], “6”」となる。図15(b)のように(固定パターンの文節情報の第1要素+3)の文節が固定パターンに係ると判定した。文節情報の要素中の並列範囲情報を利用して、並列構造の係り受けも正しく反映できる。なお、係り文節の文節情報には受け文節の文節情報

50

を追加する。受け文節を参照することでさまざまな制約をかけたり、割り当て・対応付けに利用したりできる。

【 0 0 5 1 】

係り文節集合から、図 1 3 で示したそれぞれのルールに基づいて係り受け抽出部 4 0 0 が抽出を行う (S 4 1 0)。ここで、「製品種別・製品名・細分類の抽出」では、係り文節情報からの解候補を受け文節とする係り文節集合が必要になった場合、ここから固定パターン係り受け作成クラスを呼び出している。抽出解作成部 4 1 0 がそれぞれの抽出が終了した後表記パターンによる削除や重複要素の削除を行い、抽出解とする (S 4 1 1)。本文終了か否かを判断する (S 4 1 2)。本文が終了していない場合に S 4 0 1 に戻る。本文が終了していると判断した場合に、係り受け対応・割付部 4 2 0 が抽出解から割り当て・対応付けを行う (S 4 1 3)。なお、最終的に、“記事番号”、“製品 1 [販売元、発売日、製品種別、製品名、細分類、価格]”、“製品 2 [同]”、・・・の解を表示することもできる (S 4 1 4)。

10

【 0 0 5 2 】

[6 . 売り情報の出力結果]

図 1 6 は、本発明の実施形態に係る情報抽出装置における補足説明となる語句を含めて抽出された記事の説明図である。

見出し情報は、「[ビジネス情報] 省エネタイプの自動販売機を開発 - - サンデン」である。本文情報は、「サンデンは 1 5 日、料金の安い深夜電力だけを利用して運転コストを従来機の 3 0 % に抑えた省エネタイプの自動販売機を開発したと発表した。深夜の間に缶入り飲料水を加熱・冷却、昼間は電気を使わずに飲み物の温度を適温に維持する仕組み。外部から熱が入るのを防ぐため断熱材を従来の 3 0 ミリから 5 0 ミリに厚くしたほか、商品を補充する時の庫内の温度変化を防ぐため、専用の扉を作る工夫をした。大きさは高さ 1 9 4 センチ、幅 1 1 8 センチ、奥行き 8 6 センチで、5 2 0 本を収納できる」である。

20

【 0 0 5 3 】

本発明による情報抽出装置によって抽出した結果は、見出しの特徴情報としては、「省エネタイプ」が抽出される。また、本文より抽出される特徴情報としては、「料金の安い深夜電力だけを利用して運転コストを従来機の 3 0 % に抑えた省エネタイプ」が抽出される。さらに、「売り」情報としては、「料金の安い深夜電力だけを利用して運転コストを従来機の 3 0 % に抑えた省エネタイプ」が抽出されることとなる。

30

【 0 0 5 4 】

以上の前記実施形態により本発明を説明したが、本発明の技術的範囲は実施形態に記載の範囲には限定されず、これら各実施形態に多様な変更又は改良を加えることが可能である。そして、かような変更又は改良を加えた実施の形態も本発明の技術的範囲に含まれる。このことは、特許請求の範囲及び課題を解決する手段からも明らかである。

【 図面の簡単な説明 】

【 0 0 5 5 】

【 図 1 】本発明の実施形態に係る情報抽出装置における新製品紹介記事と記事中に含まれる特徴情報の説明図である。

40

【 0 0 5 6 】

本発明の実施形態に係る情報抽出装置のハードウェア構成図である。

【 図 2 】本発明の実施形態に係る情報抽出装置における見出し中に「売り」情報を含む記事の説明図である。

【 図 3 】本発明の実施形態に係る情報抽出装置における本文中に見出しの特徴情報の補足情報を含む記事の説明図である。

【 図 4 】本発明の実施形態に係る情報抽出装置のハードウェア構成図である。

【 図 5 】本発明の実施形態に係る情報抽出装置のブロック構成図である。

【 図 6 】本発明の実施形態に係る情報抽出装置における処理フローシートである。

【 図 7 】本発明の実施形態に係る情報抽出装置における形態素解析結果である。

50

【図 8】本発明の実施形態に係る情報抽出装置における係り受け解析結果である。

【図 9】本発明の実施形態に係る情報抽出装置におけるテンプレート抽出のブロック構成図である。

【図 10】本発明の実施形態に係る情報抽出装置における分析用タグ付ききデータ例の説明図である。

【図 11】本発明の実施形態に係る情報抽出装置におけるテンプレート抽出処理のフローシートである。

【図 12】本発明の実施形態に係る情報抽出装置における係り受け解析による抽出のブロック構成図である。

【図 13】本発明の実施形態に係る情報抽出装置における固定パターンと格形式の説明図である。

10

【図 14】本発明の実施形態に係る情報抽出装置における係り受け抽出処理のフローシートである。

【図 15】本発明の実施形態に係る情報抽出装置における構文解析結果及びその文節情報の対応の説明図である。

【図 16】本発明の実施形態に係る情報抽出装置における補足説明となる語句を含めて抽出された記事の説明図である。

【符号の説明】

【 0 0 5 7 】

- | | | |
|-----|---------------|----|
| 1 | コンピュータ | 20 |
| 2 | CPU | |
| 3 | メインメモリ | |
| 4 | HDD | |
| 5 | ビデオカード | |
| 6 | マウス | |
| 7 | キーボード | |
| 8 | 光学ディスク | |
| 10 | 記事入力部 | |
| 20 | テンプレート抽出部 | |
| 30 | 係り受け抽出部 | 30 |
| 40 | タグパターンマッチング部 | |
| 50 | 記事分割部 | |
| 60 | 見出しの形態素解析部 | |
| 70 | 見出しの助詞除去部 | |
| 80 | 見出し特徴情報マッチング部 | |
| 90 | 見出し特徴情報抽出部 | |
| 100 | 本文の形態素解析部 | |
| 110 | 本文の助詞除去部 | |
| 120 | 本文特徴情報マッチング部 | |
| 130 | 本文特徴情報抽出部 | 40 |
| 140 | 本文の係り受け解析部 | |
| 150 | 補足説明の抽出部 | |
| 160 | 売り情報の出力 | |
| 210 | 記事句点分割部 | |
| 220 | Aテンプレートマッチング部 | |
| 230 | Aテンプレート抽出部 | |
| 240 | 制約チェック部 | |
| 250 | テンプレートID記憶部 | |
| 260 | Bテンプレートマッチング部 | |
| 270 | Bテンプレート抽出部 | 50 |

- 280 テンプレート製品対応部
- 310 係り受けタグパターンマッチング部
- 320 係り受けタグ分割部
- 330 見出し分析部
- 340 見出し処理部
- 350 本文句点分割部
- 360 括弧内数値判定部
- 370 文節情報作成部
- 380 固定パターン判定部
- 390 固定パターン係り受け作成部
- 400 係り受け抽出部
- 410 抽出解作成部
- 420 係り受け対応・割付部

【図1】

新製品紹介記事と記事中に含まれる特徴情報

[ビジネス情報]ニューロ制御のヒーター---三菱電機

三菱電機は、設定温度を自動的に決めるニューロ制御の石油ファンヒーター8タイプ24機種を8月21日販売する。6-16畳向けで価格は3万9000-8万9800円。

[温度設定を自動的に決める, ニューロ制御, 6-16畳向け]

[出典 毎日新聞より]

【図2】

見出し中に「売り」情報を含む記事

[雑記帳]足利銀行が視覚障害者が利用できるATM

足利銀行が、沖電気工業と共同で全国で初めて視覚障害者が利用できるATMを開発した。従来のATMに、電話の受話器のようなハンドセットと取引金額や残高がボードに点字で浮き上がる点字表示装置、点字付き操作ボタンが加わった。金額や残高が点字で表示装置にできる。同行は「社会貢献活動の一環。より良い企業市民を目指します」とアピール。

「売り」情報：視覚障害者が利用できる

[出典 毎日新聞より]

【図3】

本文中に見出しの特徴情報の補足情報を含む記事

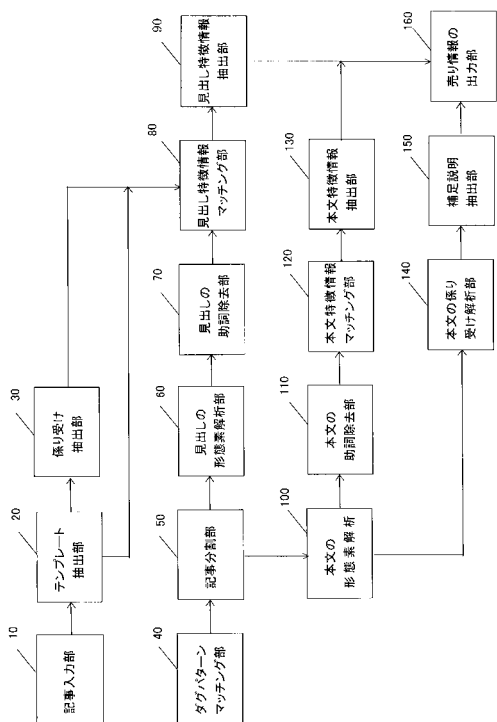
【ビジネス情報】BSチューナーを2台内蔵したテレビを発売——シャープ

シャープは、業界で初めて衛星放送チューナーを2台内蔵した29型カラーテレビ「Tツイン」を4月15日発売。22万円。衛星放送の裏番組録画が可能。衛星放送を見ながら、手持ちのVTRで衛星放送の裏番組録画が可能な同チューナーを内蔵して、別のテレビとケーブルで接続すれば、2つの衛星放送番組を同時に視聴できる。

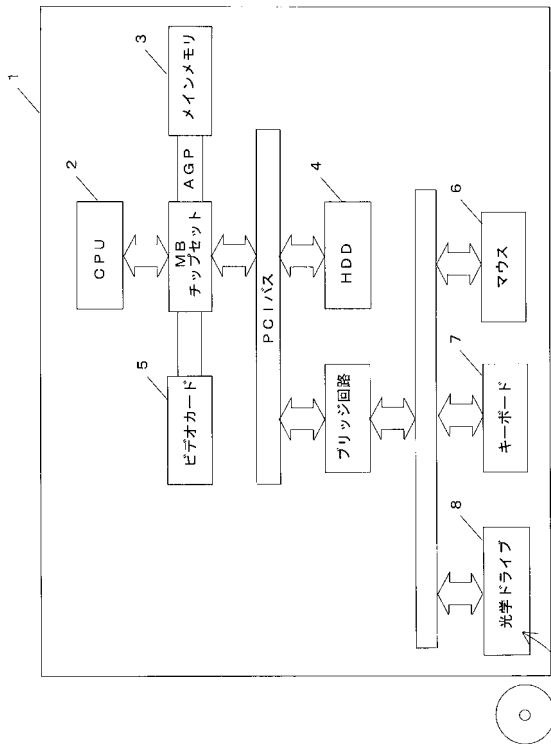
補足情報：業界で初めて

[出典 毎日新聞より]

【図5】



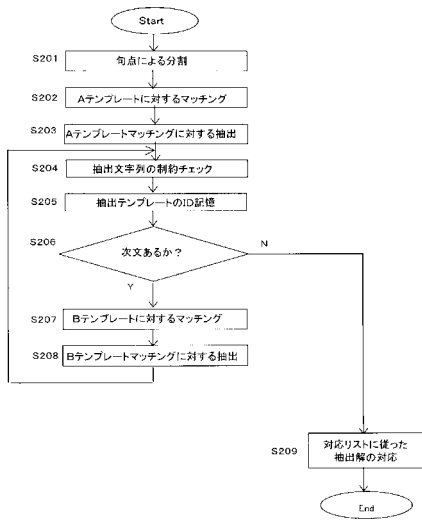
【図4】



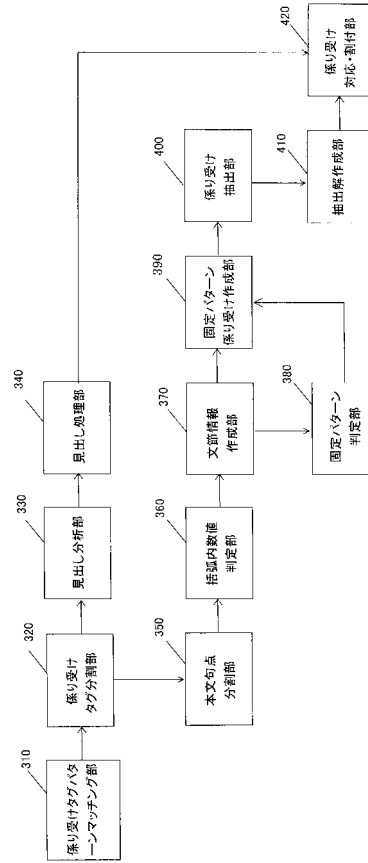
【図6】

ウガセがウガ名詞6組織名6*0*0NIL
 ...特殊1記号5*0*0NIL
 エンタープライゼスエンタープライゼスエンタープライゼス未定義語15カタカナ2*0*0NIL
 社社社社名詞6普通名詞1*0*0"漢字読み音組織名末尾代表表記社"
 ははは助詞9助詞2*0*0NIL
 ...特殊1記号2*0*0NIL
 SF SF SF 未定義語15アルファベット3*0*0NIL
 映画えいが映画名詞6普通名詞1*0*0"代表表記映画"
 的なてきなだ接尾辞14形容詞性名詞接尾辞6+形容詞21夕列基本連体形4NIL
 光線こうせん光線名詞6普通名詞1*0*0"代表表記光線"
 銃じゅう銃名詞6普通名詞1*0*0"漢字読み音代表表記銃"
 戦いくさ戦名詞6普通名詞1*0*0"漢字読み音代表表記戦"
 戦せん戦名詞6普通名詞1*0*0"漢字読み音代表表記戦"
 ををを助詞9格助詞1*0*0NIL
 模擬もぎ模擬名詞6サ変名詞2*0*0"代表表記模擬"
 体験たいけん体験名詞6サ変名詞2*0*0"代表表記体験"
 できるできる動詞2*0"可音動詞1基本形2"代表表記出来る"
 おもちやおもちやおもち動詞2*0"下音動詞1行12音便条件形14"補文代表表記思う"
 ...特殊1括弧始3*0*0NIL
 ロックろっくロック名詞6サ変名詞2*0*0"代表表記ロック"
 オンおんオン名詞6サ変名詞2*0*0"代表表記オン"
 ...特殊1括弧終4*0*0NIL
 ををを助詞9格助詞1*0*0NIL
 発売ははは発売名詞6サ変名詞2*0*0"代表表記発売"
 したしたする動詞2*0"サ変動詞16夕形8"付属動詞候補(基本)代表表記する"
 ...特殊1句点1*0*0NIL

【図 1 1】



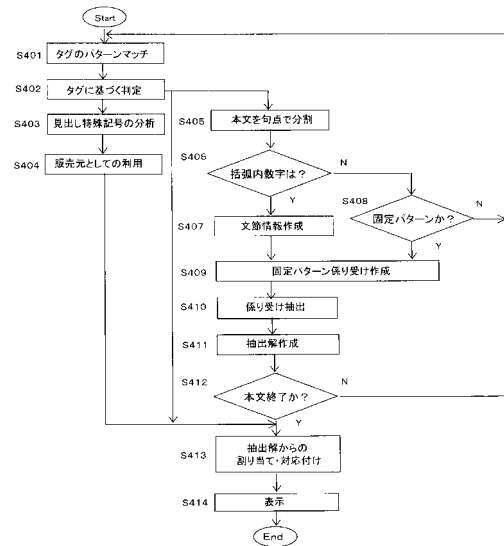
【図 1 2】



【図 1 3】

1文字目	
固定ハターン	抽出情報とみなす格形式
発売, 販売, 変更, チェンジ, 開発, 改良, 開始, 始め,	"未格, 未格" kns
発表, 商品化, 製品化, 輸入, 発刊, 発行, 出版, 創刊, 刊行	"カワ格, 隣接, 無格, ニ格" d "ワ格" kns
展開, 参入, 提携	"未格, 未格" d
変更, 強化, 強調	"カワ格, 隣接, 無格, ニ格" d
採用, 導入	"ワ格" ニ格" kns
追加, 設定, 加え	"ニ格, ワ格" kns
搭載, 装備	"名目でノ格, 用言でニ格" kns
2文字目以降	
固定ハターン	抽出情報とみなす格形式
製品数(文末)	"ノ格, 同格未格" kns
発売(文末)	"未格" kns
発売(文頭)	受け文節" kns
別売(ワノ格)	受け文節" kns
別売(ワノ格以外)	"ワ格" kns
その他	
固定ハターン	抽出情報とみなす格形式
軸格表記	"ワ格, 未格, ノ格, 隣接" kns
【目】	"同格連体, ノ格" kns
製品数, 新製品, 新商品, 新モデル(等)	"ノ格, 同格未格" kns

【図 1 4】



【図 15】

(a)

〇〇〇〇は——
 低コストの——
 店舗監視記録用ビデオ——
 「SR-L900」を——
 2日に——
 発売する。

(b)

...

2日に—— [3, "2日に", [0, 0], "価格", "体言"], "5"]
 発売する。 [0, "発売する.", "0, 0", [nil, "用言"], "6"]

【図 16】

補足説明となる語句を含めて抽出された記事

[ビジネス情報]省エネタイプの自動販売機を開発ーサンデン

サンデンは、15日、料金の安い深夜電力だけを利用して運転コストを従来機の30%に抑えた省エネタイプの自動販売機を開発したと発表した。深夜の間に缶入り飲料水を加熱・冷却、昼間は電気を使わずに飲み物の温度を適温に維持する仕組み。
 外部から熱が入るのを防ぐため断熱材を従来の30ミリから50ミリに厚くしたほか、商品を補充する時の庫内の温度変化を防ぐため、専用の扉を作る工夫をした。
 大きさは高さ194センチ、幅118センチ、奥行き86センチで、520本を収納できる。

- ・見出しの特徴情報：省エネタイプ
- ・本文より抽出される特徴情報：
料金の安い深夜電力だけを利用して運転コストを従来機の30%に抑えた省エネタイプ
- ・「売り」情報：
料金の安い深夜電力だけを利用して運転コストを従来機の30%に抑えた省エネタイプ

[出典 毎日新聞より]

フロントページの続き

- (56)参考文献 特開平03 - 135669 (JP, A)
特開2000 - 067080 (JP, A)
特開2003 - 099429 (JP, A)
特開2002 - 334091 (JP, A)
赤松順子、外4名、複数製品の紹介記事からの製品情報抽出、情報処理学会研究報告(2000 - NL - 140)、日本、社団法人情報処理学会、2000年11月22日、第2000巻、第107号、p. 61 - 68
細野公男、外2名、パターン・マッチングによる重要語の自動抽出、自然言語処理研究会資料39、日本、社団法人情報処理学会自然言語処理研究会、1983年 9月16日、p. 1 - 8
斉藤公一、外5名、数値情報をキーとした新聞記事からの情報抽出、情報処理学会研究報告(98 - NL - 125)、日本、社団法人情報処理学会、1998年 5月29日、第98巻、第48号、p. 63 - 70

- (58)調査した分野(Int.Cl., DB名)
G06F 17/30