(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2008-250887
(P2008-250887A)
(43)公開日　平成20年10月16日(2008.10.16)

| (51)Int.Cl. | | F I | | | テーマコード（参考） |
|---|---|---|---|---|---|
| G06F 17/30 | (2006.01) | G06F | 17/30 | 220Z | 5B075 |
| G06F 17/28 | (2006.01) | G06F | 17/30 | 170A | 5B091 |
| | | G06F | 17/28 | X | |

審査請求 未請求　請求項の数 8　ＯＬ　（全 24 頁）

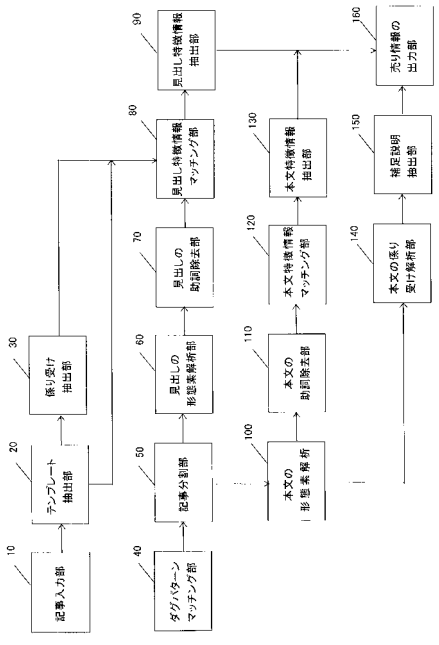
(21)出願番号　特願2007-94339(P2007-94339)
(22)出願日　平成19年3月30日(2007.3.30)

(71)出願人　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

Ｆターム(参考)　5B075 ND03　NK10　NK32　UU40
5B091 AA15　CA02　CA05

(54)【発明の名称】情報抽出装置、その方法及びプログラム

(57)【要約】　（修正有）
【課題】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。
【解決手段】記事入力手段１０と、テンプレートによる抽出手段２０と、係り受け解析による抽出手段３０と、記事情報のタグパターンマッチング手段４０と、見出しと本文の分割手段５０と、見出しの形態素解析手段６０と、見出しの助詞除去手段７０と、テンプレート又は係り受け解析により抽出された情報と見出し情報とのマッチング手段８０と、見出しの特徴情報抽出手段９０と、本文の形態素解析手段１００と、本文の助詞除去手段１１０と、見出し特徴情報と本文の文節とのマッチング手段１２０と、本文の特徴情報の抽出手段１３０と、見出し特徴情報又は本文特徴情報の売り情報出力手段１４０とを含む。
【選択図】図５



【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
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。
【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。
【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、
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、請求項３又は請求項４に記載された情報抽出装置。
【請求項６】
　前記係り受け解析抽出手段は、
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係り受け抽出手段と、
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ないし請求項４のいずれかに記載された情報抽出装置。
【請求項７】
　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ダ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。

【００１３】

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。

【００１４】

（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。

【００１５】

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。

【００１６】

（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、前記係り受け解析抽出手段は、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係り受け抽出手段と、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、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ても抽出できる。

【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ダ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。

【００２４】

（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ものではなく、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】

【００２５】

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。

【００２６】

（本発明の第１の実施形態）

［１．ハードウエア構成］

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。

【００２７】

［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、ダ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。

【００２８】
［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ダ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（Ｓ９００）。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（Ｓ１０００）。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（Ｓ１１００）本文の助詞除去部１１０が助詞の除去をする（Ｓ１２００）。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（Ｓ１３００）。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（Ｓ１４００）。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（Ｓ１５００）。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（Ｓ１６００）。売り情報の出力１６０が売り情報の出力を行う（Ｓ１７００）。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「ゃ」や「ぁ」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）抽出項目の種類の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、「＜ｃ１＞富士通ゼネラル＜／ｃ１＞は横長でワイド感のある画面を用いた＜ｋ１＞２９型衛星放送（ＢＳ）内臓テレビ＜／ｋ１＞「＜ｎ１＞ＢＳ－２９Ｍ５５＜／ｎ１＞」と「＜ｎ２＞ＢＳ－２９Ｍ５０＜／ｎ２＞」を＜ｄ１＞９月２日＜／ｄ１＞に発売する」に対して、Ｉ＝＜ｃ１，ｄ１，ｋ１，ｎ１，０，０＞－＜ｃ１，ｄ１，ｋ１，ｎ２，０，０＞となる。

【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。

｛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。

【００３４】

【数１】

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

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>２５０万円</p1>（消費税含まず）。

テンプレートは、それぞれ＜ｓ１＞と＜ｓ２＞の２種類で、＜ｓ１＞は＊、＜ｐ１＞はｐ１であるとし、頻度は、マッチした回数２０回、正解を返した回数１０回とすると、抽出する抽出情報はｓ１，ｓ２，ｓ１，ｐ１の４つであるので、以下の式が与えられる。

【００３６】

【数２】

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

【００３７】

一般にはｓ１，ｓ２，ｐ１，ｐ２あるいはｓ１，ｐ１，ｓ２，ｐ２の形をとるものが多いが、形はさまざまである。例えば、ｓ１がさらに細分化されており、ｓ１（ｓ３，ｐ３，ｓ４，ｐ４），ｓ２，ｐ２のようにｓ３，ｐ３，ｓ４，ｐ４が埋め込まれていてｓ１に対応するｐ１がもともとない場合もある。このような場合でもｓ３，ｐ３，ｓ４，ｐ４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「テンプレート作成用データの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。

【００３８】

図１１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２１０が、入力記事を区点で分割する（Ｓ２０１）。Ａテンプレートマッチング部２１０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（Ｓ２０２）。Ａテンテンプレート抽出部２３０がマッチングに対する抽出を行う（Ｓ２０３）。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２４０が制約チェックを行う（Ｓ２０４）。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２５０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（Ｓ２０５）。次の文があるか否かを判断する（Ｓ２０６）。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２６０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（Ｓ２０７）。Ｂテンプレート抽出部２７０がマッチングに対する抽出を行う（Ｓ２０８）。そして、制約チェック部２４０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（Ｓ２０４）。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２８０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（Ｓ２０９）。

【００３９】

［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１２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３０は、係り受けタグパターンマッチング部３１０、係り受けタグ分割部３２０、見出し分析部３３０、見出し処理部３４０、本文句点分割部３５０、括弧内数値判定部３６０、文節情報作成部３７０、固定パターン判定部３８０、固定パターン係り受け作成部３９０、係り受け抽出部４００、抽出解作成部４１０、係り受け対応・割付部４２０を含む。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。

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，ニ格”：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，ニ格”：ｄとする。

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＜ーノ格”：ｋｎｓとする。固定パターンが、「採用」、「導入」について、“ヲ格＜ーニ格”：ｋｎｓとする。固定パターンが、「追加」、「設定」、「加え」について、“ニ格，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，用言でニ格”：ｋｎｓとする。

【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、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。

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

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，ニ格，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、ニ格、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“ニ格、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、ニ格”で係る文節。（３）“採用、導入、強化、強調”に“ニ格”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“ニ格”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、それを受ける文節。（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”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」、「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。

【００４８】

【表２】

「((約|同|計)？([，、一二三四五六七八九十百千万億〇１２３４５６７８９０・]＋円)＋(—|～|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)？)＋(—)([，、一二三四五六七八九十百千万億〇１２３４５６７８９０・]＋円)＋(程度|台)？)」

【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。

【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。

【００５１】

係り文節集合から、図１３で示したそれぞれのルールに基づいて係り受け抽出部４００が抽出を行う（Ｓ４１０）。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４１０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（Ｓ４１１）。本文終了か否かを判断する（Ｓ４１２）。本文が終了していない場合にＳ４０１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４２０が抽出解から割り当て・対応付けを行う（Ｓ４１３）。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（Ｓ４１４）。

【００５２】

［６．売り情報の出力結果］

図１６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」である。

【００５３】

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３０％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３０％に抑えた省エネタイプ」が抽出されることとなる。

【００５４】

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なことである。

【図面の簡単な説明】

【００５５】

【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。

【００５６】

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。

【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。

【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】

【００５７】

１　コンピュータ
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
４０　ダ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
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
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。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。

「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

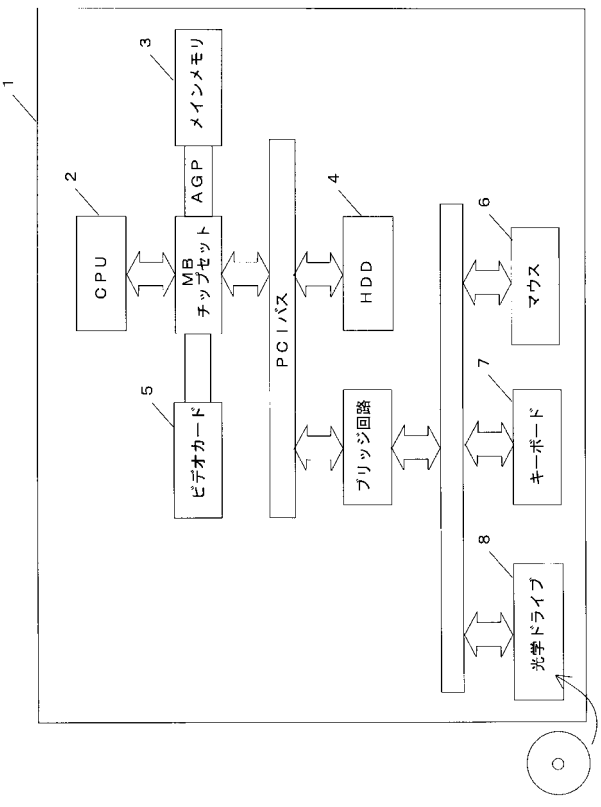
［ビジネス情報］ＢＳチューナーを２台内蔵したテレビを発売－－シャープ

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「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。

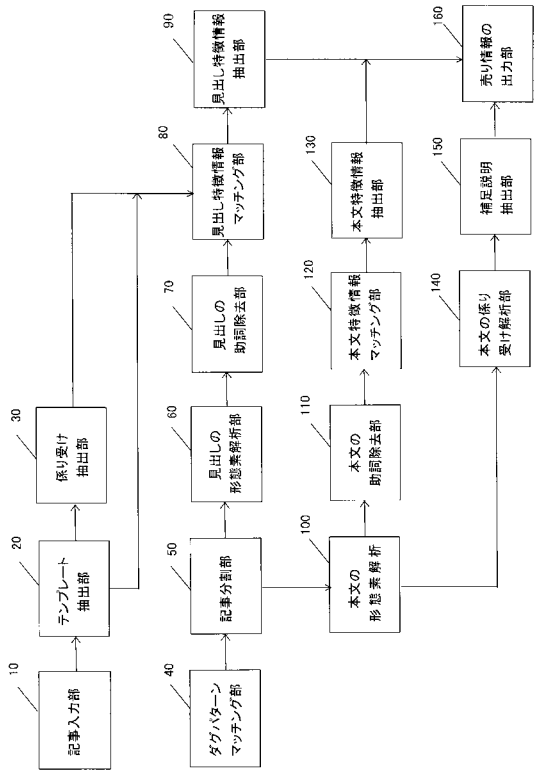
補足情報：業界で初めて

［出典　毎日新聞より］

【図４】



【図５】



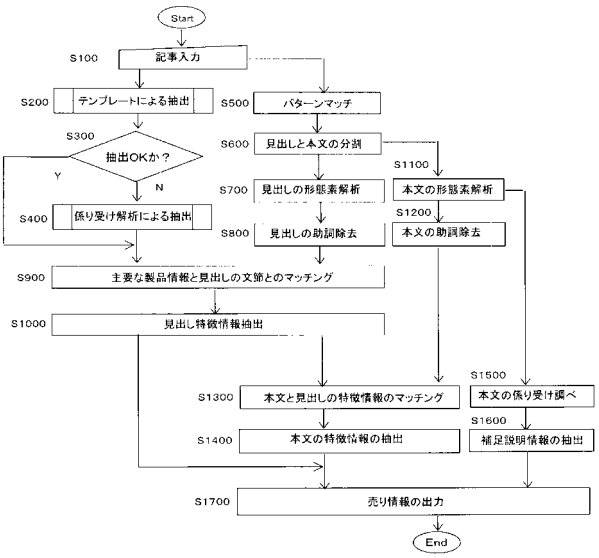
【図６】

```
セガ せが セガ 名詞 6 組織名 6 * 0 * 0 NIL
・ ・ ・ 特殊 1 記号 5 * 0 * 0 NIL
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 * 0 NIL
社 しゃ 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 副助詞 2 * 0 * 0 NIL
、 、 、 特殊 1 読点 2 * 0 * 0 NIL
ＳＦ ＳＦ ＳＦ 未定義語 15 アルファベット 3 * 0 * 0 NIL
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 的だ 接尾辞 14 形容詞性名詞接尾辞 6 ナ形容詞 21 ダ列基本連体形 4 NIL
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:訓 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 * 0 NIL
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 できる 動詞 2 * 0 母音動詞 1 基本形 2 "代表表記:出来る"
おもちゃ おもちゃ おもう 動詞 2 * 0 子音動詞ワ行 12 音便条件形 14 "補文ト 代表表記:思う"
「 「 「 特殊 1 括弧始 3 * 0 * 0 NIL
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 * 0 NIL
を を を 助詞 9 格助詞 1 * 0 * 0 NIL
発売 はつばい 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 タ形 8 "付属動詞候補 (基本) 代表表記:する"
。 。 。 特殊 1 句点 1 * 0 * 0 NIL
```

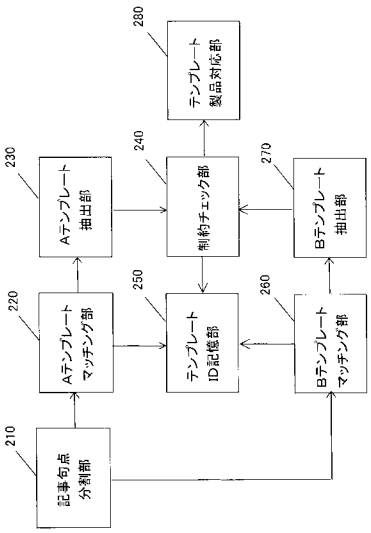
【図７】

```
((6 (type:D) ((発売 はつばい 発売 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:発売" (代表表
記:発売 漢字 かな漢字 名詞相当語 サ変 サ変動詞 自立 タグ単位始 文節始 ))(した した す
る 動詞 2 * 0 サ変動詞 16 タ形 8 "付属動詞候補 (基本) 代表表記:する" (付属動詞候補 (基
本) 代表表記:する 表現文末 とタ系連用テ形複合辞 かな漢字 ひらがな 活用語 付属 ))(。 。 。
特殊 1 句点 1 * 0 * 0 NIL (文末 英記号 記号 付属 ))) (文末 サ変 サ変スル 時制:過去 句
点 用言:動 レベル:C 区切:5-5 ID: (文末) RID:112 提題受:30) NIL)
  ((5 (type:D) ((「 「 「 特殊 1 括弧始 3 * 0 * 0 NIL (記英数カ 英記号 記号 括弧始 括
弧 接頭 非独立接頭辞 タグ単位始 文節始))(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" (代表表記:ロック 記英数カ カタカナ 名詞相当語 サ変 自立 固有キー))(オ
ン おん オン 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:オン" (代表表記:オン記英数カ カタカ
ナ 名詞相当語 サ変 自立 複合← タグ単位始 固有キー))(」 」 」 特殊 1 括弧終 4 * 0 * 0 NIL
(記英数カ 英記号 記号 括弧終 括弧 述語区切 付属 ))(を を を 助詞 9 格助詞 1 * 0 * 0 NIL
(かな漢字 ひらがな 付属 ))) (サ変 ヲ 括弧始 括弧終 助詞 体言 係:ヲ格 区切:0-0 RID:1099
格要素 連用要素) NIL)
    ((4 (type:D) ((おもちゃ おもちゃ おもう 動詞 2 * 0 子音動詞ワ行 12 音便条件形 14
"補文ト 代表表記:思う" (補文ト 代表表記:思う かな漢字 ひらがな 活用語 自立 タグ単位始 文
節始))) (補文ト 用言:動 係:NONE RID:1458) NIL))
    ((3 (type:D) ((模擬 もぎ 模擬 名詞 6 サ変名詞 2 * 0 * 0 "代表表記:模擬" (代表
表記:模擬 漢字 かな漢字 名詞相当語 サ変 自立 タグ単位始 文節始 ))(体験 たいけん 体験 名
詞 6 サ変名詞 2 * 0 * 0 "代表表記:体験" (代表表記:体験 漢字 かな漢字 名詞相当語 サ変 サ
変動詞 自立 複合← タグ単位始 ))(できる できる できる 動詞 2 * 0 母音動詞 1 基本形 2 "
代表表記:出来る" (代表表記:出来る かな漢字 ひらがな 活用語 付属 ))) (サ変 可能表現 態:
可能 ～できる 連体修飾 用言:動 タグ単位受無視 係:連格 レベル:B 区切:0-5 ID: (動詞連体)
RID:699 連体並列条件) NIL)
      ((2 (type:D) ((光線 こうせん 光線 名詞 6 普通名詞 1 * 0 * 0 "代表表記:光線"
(代表表記:光線 漢字 かな漢字 名詞相当語 自立 タグ単位始 文節始 ))(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" (漢字読み:音 代表表記:銃 漢字 かな漢字 名詞
相当語 自立 複合← タグ単位始 ))(戦 せん 戦 名詞 6 普通名詞 1 * 0 * 0 "漢字読み:音 代
表表記:戦" (漢字読み:音 代表表記:戦 品曖 音訓解消 ALT-戦-いくさ-戦-6-1-0-0-"漢字読み:
訓 代表表記:戦" 品曖-普通名詞 品曖-その他 漢字 かな漢字 名詞相当語 自立 複合← タグ単
位始 ))(を を を 助詞 9 格助詞 1 * 0 * 0 NIL (かな漢字 ひらがな 付属 ))) (ヲ 助詞 体
言 係:ヲ格 区切:0-0 RID:1099 格要素 連用要素) NIL)
        ((1 (type:D) ((ＳＦ ＳＦ ＳＦ 名詞 6 組織名 6 * 0 * 0 NIL(品詞変更:ＳＦ
-ＳＦ-ＳＦ-15-3-0-0 品曖 品曖-アルファベット 品曖-組織名 記英数カ 英記号 名詞相当語 自
立 タグ単位始 文節始 固有キー))(映画 えいが 映画 名詞 6 普通名詞 1 * 0 * 0 "代表表記:
映画" (代表表記:映画 漢字 かな漢字 名詞相当語 自立 複合← タグ単位始 ))(的な てきな 的
だ 接尾辞 14 形容詞性名詞接尾辞 6 ナ形容詞 21 ダ列基本連体形 4 NIL (類似計算:的 かな漢
字 活用語 用体変化 付属 非独立有意味接尾辞 ))) (連体修飾 用言:形 タグ単位受無視 係:連
格 レベル:B- 区切:0-5 ID: (形判連体) RID:761 連体並列条件) NIL)))))
  ((0 (type:D) ((セガ せが セガ 名詞 6 組織名 6 * 0 * 0 NIL (文頭 記英数カ カタカナ
名詞相当語 自立 タグ単位始 文節始 固有キー))(・ ・ ・ 特殊 1 記号 5 * 0 * 0 NIL (記英
数カ 付属 複合← ))(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 名詞 6 サ
変名詞 2 * 0 * 0 NIL (品詞変更: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-15-2-0-0 品曖-カタカナ 記英数カ カタカナ 名詞相当語 サ変 自立 複合← タグ単位始 固有
キー))(社 しゃ 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" (漢字
読み:音 組織名末尾 代表表記:社 漢字 かな漢字 名詞相当語 自立 複合← タグ単位始 ))(は は
は 助詞 9 副助詞 2 * 0 * 0 NIL (かな漢字 ひらがな 付属 ))(、 、 、 特殊 1 読点 2 * 0 *
 0 NIL (英記号 記号 述語区切 付属 ))) (文頭 サ変 組織名 ハ 読点 助詞 体言 係:未格 提
題 区切:3-5 RID:1252 格要素 連用要素) NIL)))
```

【図８】



【図９】




【図１０】

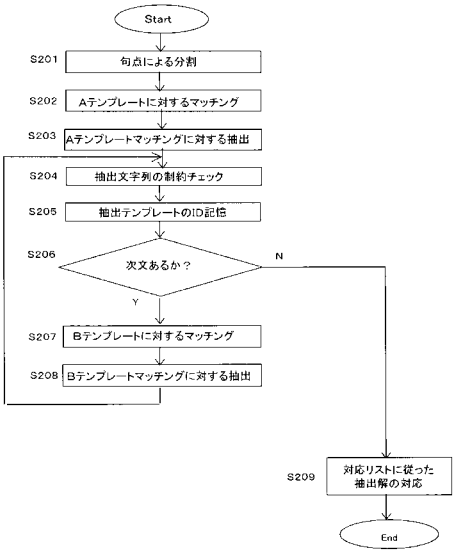
分析用タグ付きデータ例

〈ｃ１〉富士通ゼネラル〈／ｃ１〉は横長でワイド感のある画面を用いた〈ｋ１〉２９型衛星放送（ＢＳ）内臓テレビ〈／ｋ１〉「〈ｎ１〉ＢＳ－２９Ｍ５５〈／ｎ１〉」と「〈ｎ２〉ＢＳ－２９Ｍ５０〈／ｎ２〉」を〈ｄ１〉９月２日〈／ｄ１〉に発売する。

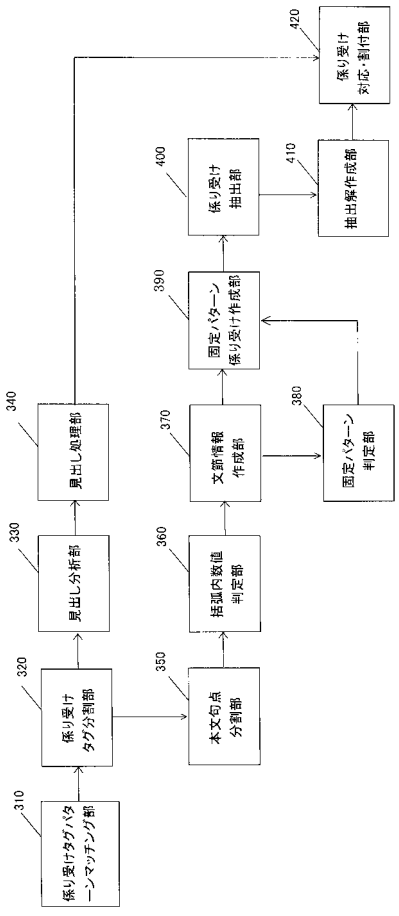
Ｉ＝〈ｃ１，ｄ１，ｋ１，ｎ１，０，０〉－〈ｃ１，ｄ１，ｋ１，ｎ２，０，０〉

［出典　毎日新聞より］

【図１１】



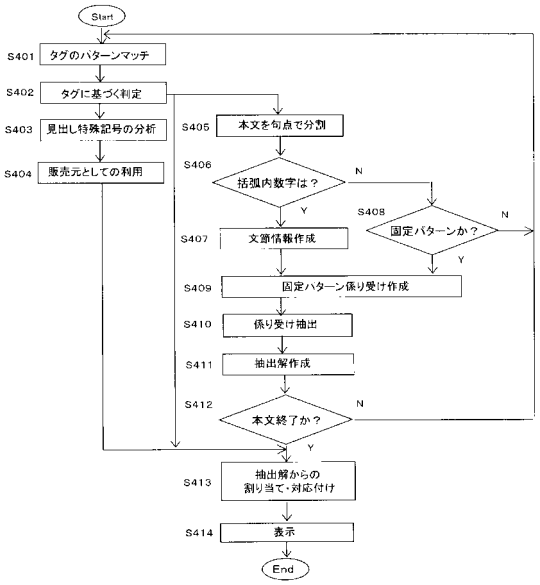

【図１２】



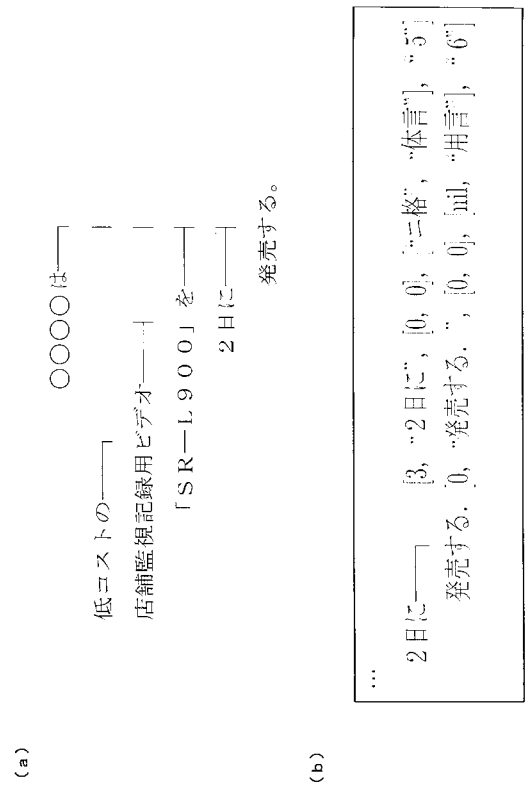
【図１３】

| 1文目 | |
|---|---|
| 固定パターン | 抽出情報となる文格形式 |
| 発売,販売,売り,チェンジ,<br>開発,改良,開始,始め,<br>発表,商品化,製品化,輸入,<br>発刊,発行,出版,創刊,刊行 | "未格,ガ格:c"<br>"カラ格,隣接,無格,ニ格:d"<br>"ヲ格":kns |
| 展開,参入,提携 | "未格,ガ格:c"<br>"カラ格,隣接,無格,ニ格:d" |
| 変更,強化,強調 | "ヲ格ゞ一ノ格":kns |
| 採用,導入 | "ヲ格ゞ一ニ格":kns |
| 追加,設定,加え | "ニ格,ヲ格":kns |
| 搭載,装備 | "依高ゞノ格,用言ゞニ格":kns |
| 2文目以降 | |
| 固定パターン | 抽出情報となる文格形式 |
| 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 |
| 別売り(ノ格) | "受け文節:kns |
| 別売り(ノ格以外) | "デ格":kns |
| その他 | |
| 固定パターン | 抽出情報となる文格形式 |
| [価格表記] | "ガ格,未格,デ格,隣接":kns |
| [同] | "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１４】




【図１５】




【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－－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］