新聞記事を対象とした企業動向に関する事象構造の抽出

桑畑和佳子, 橋本三奈子 *1　　木田敦子 *2　　落谷亮, 西野文人 *3

*1 富士通　　*2 計量計画研究所　　*3 富士通研究所

1 はじめに

新聞記事を検索する場合、単なるキーワード検索よりも、検索対象のテキストや問い合わせを構造化した上で検索した方が、検索精度が高まることが報告されている[1]。検索対象のテキストをあらかじめ構造化しておく方法としては、形態素解析に基づいてボトムアップに解析する方法と、字面のパターン照合に基づいてトップダウンに解析する方法とが考えられるが、新聞記事のように、固有名詞が多く、また、半定型的であるテキストを対象とする場合には、後者の方法が有効であると考えられている[2]。

そこで我々は、後者のアプローチに基づいて、新聞記事を対象とする構造化テキスト検索システムを構築した。このシステムでは、構造化処理の第一段階として、新聞記事の一文目の文末表現に着目して「事象」を判定する。ここで事象とは、例えば「製品販売」や「組織合併」のような出来事のことである。

事象を判定することによって、事象ごとに付随する属性(例えば、製品販売の場合には、販売元や発売日などの情報)を判断して後続の解析処理を進めるため、文末表現からの事象の判定は、構造化処理全体の成否を左右する非常に重要な処理である。

字面のパターン照合に基づく構造化は従来から数多く試みられているが[3]、それらのほとんどは、事象が正しく判定されることを前提として、属性値の抽出精度の方を主眼として議論がなされており、事象そのものを判定する部分に関しては詳細な検討は行われていなかった。これに対し、本稿では、文末表現からの事象判定に着目して、「行為語」が述語表現を形成する規則の詳細な検討に基づいて事象の判定方法を提案するとともに、事象の判定精度に関する実験結果について述べる[1]。

2 構造化テキスト検索システム

最初に、構造化テキスト検索システムについて簡単に紹介する。

単なるキーワード検索には、意図しない文字列照合が起きるという問題がある。例えば、富士通の製品販売記事の検索を従来の単なるキーワード検索で行おうとすると、「富士通」と「販売」をAND検索で実行した結果、富士通が主体とって何かを販売する記事以外に、他社の販売記事中に富士通の名前が出ていものまで検索されてしまう。このような時は、「富士通」ではなく、「富士通は」「富士通が」などとキーワードを指定しなおせばいくらか絞り込むことも可能であるが、「富士通は」に対応しない「販売」がたまたま同記事中に出現しているようなものまでは省けない。

他方、富士通の製品販売記事の検索を構造化テキスト検索システムで行う時は、「製品販売記事」のうち、<販売組織名>が「富士通」であるもの、と指定した検索ができるため、確実に目的の記事だけが検索される。こういった記事種別選択と、属性を指定したキーワード検索が可能になるのは、あらかじめ製品販売記事を選別し、製品販売という事象の属性情報である、<販売組織名>、<発売日付>、<製品種>、<販売価格>、といった属性情報を抽出し、タグ付けしているからである。

構造化テキスト検索システムは、検索入力指定にだけではなく、結果表示にも特徴がある。製品販売記事の場合は、「どこが」「いつ」「何を」「いくらで」販売するのかが一覧できれば、どのような記事が検索されてきたかが捉えやすくなる。そこで、構造化テキストを利用して、検索結果として記事の要点となる事象の属性値のみを取り出して表示している。例として<価格>の範囲と<製品種>とを指定した「30万円以下のパソコンの販売記事」という条件で検索した時の検索結果表示を図1に示す。

このように、構造化テキスト検索システムは、あらかじめテキストから事象構造を抽出して構造化し、構造化した情報を利用して検索処理を行うことにより、冗長な検索を排除し、欲しい情報だけを確実に検索できるシステムであり、また、検索結果の提示を効率良く行うことができるシステムである。

3 事象構造の抽出

本章では、テキストを構造化するために行う事象構造の抽出について述べる。

一般に新聞記事では、一文目にその記事で扱う事象全体の要約が述べられる。我々は、企業動向に関する事象の記事の場合、その一文目の文末の述語表現にその事象に特徴的な「行為語」が現れやすいことに着目した。例えば、「製品販売」記事であればその文末に「販売」「発売」という行為語が現れる。また、「組織合併」記事であればその文末に「合併」という行為語が現れる。

また、一文目が半定型的な形である点にも注目した。例えば、表1に示すように、「販売する」「合併する」という述語表現の文に、「製品販売」や「組織合併」といった事象の属性である<組織情報>や<日付>は半定型的な形で述べられる。

このように一文目が半定型的であることを利用して、形態素解析などを使わず字面によるパターン照合を行い、新聞記事テキストからの事象構造の抽出を行った。事象構造抽出処理の第一段階として、一文目の文末表現を解析し、

[1]本研究開発は、IPA創造的ソフトウェア育成事業の支援によるものです。

注釈情報一覧表（製品販売）

全記事数 ：9 異なる記事数 ：9 表示記事数 ：1−9

組織名	製品種	価　格	発売日	
米ＩＢＭ	世界初の「ＩＢＭ・アップル互換」パソコン			米ＩＢＭ、アップル互換
富士通	値段を従来機の半分に下げたデスクトップ（机上）型パソコン	200,000〜299,999	1993/02/29	低価格パソコン、富士通
ＮＥＣ	価格性能比を既存機種の約三倍に引き上げたパソコン	200,000〜299,999	1993/02/05	パソコン、ＮＥＣも低価
ＮＥＣ	次世代パソコン		1993/07/27	次世代パソコン、ＮＥＣ
松下電器産業	高速処理ノート型パソコン	278,000	1990/11/20	松下電器、薄いノート型
富士通	ディスプレー一体型などを加えたパソコン	268,000	1991/11/05	富士通、ビジネス機能拡
富士通	うえで、三十万円を切る低価格に抑えたカラーノート型パソコン	298,000	1994/06/01	カラーノート型パソコン
米ＩＢＭ	パソコン			
富士通	ＩＢＭ互換パソコン	178,000		ＩＢＭ互換機発売、富士

全記事数 ：9 異なる記事数 ：9 表示記事数 ：1−9

図 1: 構造化テキストの検索結果

表 1: 事象別の定型文例

＜組織情報＞は、＜日付＞、＜製品＞を販売する。
＜組織情報＞は、＜製品＞を＜日付＞販売する。
＜組織情報＞と＜組織情報＞は、＜日付＞合併する。
＜組織情報＞は、＜組織情報＞と＜日付＞合併する。

その記事の事象を判定する。そして、第二段階として、第一段階の事象判定結果に基づき、事象に付随する属性とその出現パターンを予測し、一文目より事象の属性値を抽出する。一文目の解析だけで予測した属性値が全て見出せなかった時は、第三段階として、二文目以降より属性値を抽出する。

以下、事象構造の抽出処理を進めるにあたって非常に重要である第一段階の事象判定の方法についての詳細を述べる[2]

3.1 事象判定のための述語表現の照合規則

字面によるパターン照合解析で事象を判定するためには、行為語がどういった述語表現を形成するかを照合規則として用意しなければならない。そこで、日経新聞 1990 年から 1994 年まで [5]、5 年分の記事から「販売」「発売」「合併」という事象行為語が形成する述語表現を収集した。すると、「販売」という行為語については、「販売する」「販売することを決めた」「販売を開始する」「販売を発表した」といったようにさまざまな述語表現のバリエーションが見つかった。この時、「販売」と同じ事象の行為語である「発売」についても「販売」とほとんど同じ述語表現が見つかった。さらに、「販売」とは異なる事象の行為語である「合併」についても、「合併する」「合併することを決めた」「合併を発表した」といったように、行為語以外の部分では「販売」「発売」とほとんど変わらない述語表現が集まった。よって、「製品販売」と「組織合併」のような事象の異なりは、行為語の異なりとして表現され、述語表現の形成パターンは共通のものになっているという見通しがたった。

収集した述語表現の分析をさらに進めた結果、行為語が「販売」「発売」の場合には、述語表現を形成する形式は、三層に分けて捉えられることがわかった。

表 2: 述語表現形式の三層構造

第一層	第二層	第三層	形成例
行為語	開始語無	発表語無	販売する
(用言型)		発表語有	販売すると発表した
	開始語有	発表語無	販売し始めた
		発表語有	販売し始めたと発表した
行為語	開始語無	発表語無	＊体言型単独の述語形はなし
(体言型)		発表語有	〜の販売を発表した
	開始語有	発表語無	〜の販売を開始する
		発表語有	〜の販売を開始すると発表した

さまざまな述語表現は、行為語、開始語、発表語の三層で形成される。まず、第一層の行為語は、「〜を／〜に／〜で＜行為語＞」のように、学校文法でいう連用修飾を受ける用言型と、「〜の＜行為語＞」のように、連体修飾を受ける体言型とに分けられる。次に第二層の開始語には、「開始する」「始める」「し始める」のような述語があり、この第二層の述語は出現する場合としない場合とがある。最後に、第三層の発表語（「発表する」「明らかにする」など）をとるかとらないかというバリエーションがある。これらの組合せで、さまざまな述語表現が形成されている。

そこで、事象判定するためのパターン照合規則として、形成要素の語彙リストと、各形成段階の形成パターンのリストを用意した。以下、我々が作成した規則を示す。

最初に、行為語の語彙リストを事象別に表 2 に列挙する。

上にあげた第一層の行為語は、「販売する」のような単純な形式だけでなく、表 4 に示すような述語表現形成パターンで、述語表現を形成する。

[2]第二段階、第三段階の処理についての詳細は木田 [4] で述べる。

表 3: 事象別行為語の語彙リスト

製品販売	販売, 発売, 通信販売, 受注販売, セット販売, 本格販売, 実験販売, 試験販売, サンプル販売, 輸入販売
組織合併	合併, 吸収合併, 統合合併, 吸収・合併, 統合・合併, 救済合併, 対等合併, 編入合併, 買収合併, 買収・合併

表 4: ＜行為語＞の述語表現形成パターン

(A)	＜行為語＞＋＜スル動詞＞
例:	販売＋する
(B)	＜行為語＞＋ヲ＋＜行ウ動詞＞
例:	販売＋ヲ＋行う
(C)	＜行為語＞＋＜交渉中表現＞
例:	販売＋する交渉を進めている

表 4 中の各 <> の表現と第二層の開始語の語彙リストを表 5 に示す。

表 5: ＜行為語＞述語及び開始語の語彙リスト

＜スル動詞＞	する, した, される, された
＜行ウ動詞＞	行う, 行なう, 行った, 行なった
＜交渉中表現＞	する交渉を進めている, することで交渉している, 業務を新たに認めていく
＜開始語＞	開始する, 開始した, 開始される, 開始された, 始める, 始めた, し始める, し始めた, 始まる, 始まった

第三層の発表語が述語表現を形成するパターンを大きく捉えると、表 6 に示す 3 パターンがある [3]。体言型は必ず発表述語との組合せで述語表現が形成される。用言型は、体言化表現によって体言化し発表述語と組合わさるか、用言型のまま発表述語と組合わさるか、の二通りがある。

最後に、表 6 中に <> で示した、＜体言化表現＞と＜発表語＞の語彙リストを表 7 に列挙する。

以上、これまでにあげた語彙リストと形成パターンを事象判定のための照合規則とした。この規則を展開することによって、新聞記事の一文目の文末表現と照合し、事象を判定する[4]。

なお、行為語が「合併」の場合には、「合併する」という動詞が瞬間動詞に属するものであるため、「＜行為語＞＋（ヲ／モ＋）＜開始表現＞」という述語表現は有り得ない。しかし、述語の組合せを層として捉え、層ごとに述語形成パターンを作成したため、組織合併事象のために新たに照合規則を作り直す必要はなく、製品販売事象の規則を用いて、簡単に作ることができる。

また、合併事象の場合には、行為語と発表語の間の第二層に、「合意する」という表現が現れる。この表現は複数主体を主語とする述語に特有のものであるため、合意語の語彙リストと述語形成パターンを作成しておけば、さらに別の「業務提携」などの事象にも共通に利用することができる。

このように、新聞記事にさまざまに出現する述語表現

[3]実際のパターン数は、読点や発表日付表現の有無、出現位置の差異によってさらに増える。

[4]実際に規則化した際は、表 4 の各語、及び表 7 の＜発表語＞の各語は、テンスによって、あるいは、能動形か受身形かによって、事象を構成する必須要素の定型パターンが決まってくるため、さらに場合分けして規則化した。

表 6: 発表語の述語表現形成パターン

(A)	＜行為語・体言型＞＋ヲ／ガ＋＜発表語＞
例:	（～の）販売＋ヲ＋発表した。
(B)	＜行為語・用言型＞＋＜体言化表現＞＋ヲ／ガ＋＜発表語＞
例:	（～を）販売すること＋ヲ＋発表した。
(C)	＜行為語・用言型＞＋ト＋＜発表語＞
例:	（～を）販売する＋ト＋発表した。

表 7: ＜体言化表現＞と＜発表語＞の語彙リスト

＜体言化表現＞	こと, 考え, 計画, 方針
＜発表語＞	発表した, 明らかにした, 決めた, 決定した, 明らかになった, ことになった, 計画だ, 考えだ

を、行為語、開始語、合意語、発表語のように、層として捉えて、それぞれの層ごとに、語彙リストと述語形成パターンを定義し組み合わせることによって、効率的で洩れのない照合規則を作成することができ、字面によるパターン照合解析において、事象の判定を正確に行なうことができるようになった。

4 実験と評価

本章では、テキスト構造化処理の事象判定精度に関する実験と評価について述べる。

事象判定の精度を調べるために、実験対象テキストとして、あらかじめ見出しか記事本文に「発売」または「合併」というキーワードを含む新聞記事テキストを用意した。日経新聞 1990 年から 1994 年の記事を分析して照合規則を作ったので、実験は新たに日経新聞 1996 年 [5] の 9 月ひと月分の記事より、「発売」を含む 504 記事と、「合併」を含む 210 記事とを用意した。キーワード文字列検索には冗長性という問題があることは第 2 章で述べた通りである。用意したテキストには、製品販売、合併という事象を主として扱っている記事と、主ではない記事とが含まれている。次に示すような、主ではない記事が用意したテキスト中、「発売」に関して 230 記事、「合併」に関して 177 記事あった。

<headline>ミネベア社長荻野五郎氏――工場歩き現場に精通、業績に自信（リーダーの研究）</headline><body><p>積極的なM＆A（企業の合併・買収）戦略を貫いた創業者一族の故高橋高見ミネベア社長の後を受けて十年。社長の荻野五郎は半導体子会社売却など本回帰を打ち出す中で、収益を回復軌道に乗せた。（以下略）

用意したテキストに対して一文目の述語表現の照合に基づく事象判定を行った。また、事象判定の精度と比較するため、「見出しにキーワードを含む」という条件と、「一文目にキーワードを含む」という条件で検索した場合の検索精度も調べた。その結果を表 8 にまとめて示す。

キーワード検索の場合、「見出し」や「一文目」に範囲を限定することによって適合率を上げられることは実験の結果よりわかったが、事象判定による適合率の高さには及ばない。いわゆる冗長な検索結果となるものを排除するの

表 8: 実験結果

	A. 事象	B. 見出し	C. 一文目	D. 全体
販売再現率(%)	48.5	30.0	52.6	100.0
販売適合率(%)	100.0	80.2	78.5	53.6
合併再現率(%)	71.9	87.5	93.8	100.0
合併適合率(%)	100.0	40.6	38.5	15.2

A. 事象　：　一文目の述語表現の照合に基づく事象判定
B. 見出し　：　見出しに範囲を限定したキーワード検索
C. 一文目　：　一文目に範囲を限定したキーワード検索
D. 全体　：　単なるキーワード検索

に、述語表現の照合に基づく事象判定方法は効果が高いことがわかる。

しかしながら、事象判定では再現率が低い。つまり、検索もれという問題が起きているということである。合併事象に比べ、製品販売事象の「A. 事象」及び、「B. 見出し」の再現率が低いのは、以下の記事のような定型の製品紹介記事の判定に数多く失敗しているためである。

<headline> 歯ごたえある太めん、即席めん――ハウス食品（ニューフェース）</headline><body><p> ◇即席めん◇ハウス食品（06・788・1231）の「ほんしこ」＝写真 《ポイント》生めんのような歯ごたえのある太めんを採用した袋入り即席めん。豚骨スープをベースにしたしょうゆ味とみそ味がある。東日本地区で販売する。</ /p><p> 《価格・発売時期》百円。十月十六日。</ /p>

このような定型記事は、照合規則の追加で簡単に対応可能である。製品販売の事象判定に失敗したうちの 6 割がこのタイプの記事であったため、上記のような定型の記事を解析する照合規則を追加することで、製品販売の事象判定の再現率は、76.7 % にまで上げられる。

さらに、事象判定の再現率が低い原因を調べたところ、とるべき記事の判定に失敗した原因は大きく分けて二つあった。一つは、「行為語」の語彙リストのもれによるものであり、もう一つは、照合規則のうちの形成パターンリストが不十分であったことによるものである。

例えば、製品販売記事の場合、見出しでは「発売」といっているが、一文目には「売り出した」「投入した」という行為語を用いているため、事象判定に失敗したものがあった。ただし、これも「行為語」に「売り出した」「投入した」を加えることですぐに対処可能である。

また、照合規則は文末表現に限定して作成したため、次に示す記事のように、一文目が複文になっていて、先行句の句末に行為語を含む述語表現が現れるようなものの判定に失敗している。

<headline>東京と福島の縫製業、「総合型」厚年基金が合併――加入者減少、財政破たん防止。</ /headline><body><p> 都内を中心とする縫製業者でつくる厚生年金基金と、福島県のアパレル業者の基金が合併、十三日に新基金として再発足する。（以下略）

複文に対処するための照合規則の追加は、今後対処すべき課題である。

5 おわりに

本稿では、構造化テキスト検索システムにおける文末表現からの事象判定処理に着目して、「行為語」が述語表現を形成する規則の詳細な検討に基づいて事象の判定方法を提案するとともに、「製品販売」と「組織合併」を例にとって、事象の判定精度に関する実験結果を示した。

「製品販売」や「組織合併」という事象を扱った新聞記事の場合、一文目の述語表現を照合解析することで、正しく事象を判定し、単なるキーワード検索に比べ冗長性を排除する効果が高いことを示した。しかし、照合規則が十分でないと検索もれが生じるという問題があった。他の事象に対して照合規則を拡張する際は、検索もれへの十分な検討が必要になる。

また、実際に他の事象へ照合規則を拡張するにあたっては、照合解析による方法が適用可能である事象かどれだけあるかを調べる必要がある。文末の述語表現部分に事象特有の特徴がある事象であれば、この方法は適用できる。照合規則は三層に分けて作成したため、他の事象に拡張していく際は、「行為語」部分の追加や、形成パターンの若干の追加だけで対応可能である事象も少なくないだろうと予測している。例えば、「製品販売」にほとんど含まれるような「製品出荷」という事象の場合は、行為語に「出荷, サンプル出荷, 再出荷」を加えるだけで適用可能であることが現在までの予備調査で確認できている。

他の事象へ照合解析を適用することの検討はまだ着手したばかりであり、今後の課題である。

謝辞　実験用のテキストを用意するツールを提供して下さった富士通研究所の小川知也氏、並びに、本稿をまとめるにあたって貴重な意見を下さった計量計画研究所の乾裕子氏に感謝申し上げます。

参考文献

[1] 岸本行生, 須之内美幸, 塚田康博, 千葉滋, 石川徹也: テキストの構造化に基づく検索システム, 情報処理学会論文誌, vol. 35, No. 5, pp.908-916(1994).

[2] 西野文人, 落谷亮, 木田敦子, 乾裕子, 桑畑和佳子, 橋本三奈子: トップダウンなパターン解析に基づく情報抽出, 情報処理学会第 124 回自然言語処理研究会 (1998).

[3] 松尾比呂志, 木本晴夫: 抽出パターンからの階層的照合に基づく日本語テキストからの内容抽出法情報処理学会論文誌, vol. 36, No. 8, pp.1838-1844(1995).

[4] 木田敦子, 乾裕子, 桑畑和佳子, 橋本三奈子, 落谷亮, 西野文人: 情報抽出のための新聞記事テキスト分析, 言語処理学会第 4 回年次大会 (1998).

[5] 日経全文記事データベース CD-ROM 1990 年版,1991 年版,1992 年版, 1993 年版,1994 年版,1996 年版.