情報抽出のための文末表現分析

木田 敦子[1]　乾 裕子[1,2]　落谷 亮[3]　西野 文人[3]

1 計量計画研究所　2 九州工業大学大学院　3 富士通研究所

{akida, hinui}@ibs.or.jp　h_inui@pluto.ai.kyutech.ac.jp　{ochi, nisino}@flab.fujitsu.co.jp

1 はじめに

我々は、新聞記事から事象および事物を定義する表現を抽出することで、知識データベースの自動作成を行う研究を進めている。

これまでに我々は、新聞記事の表層パターンに着目することで企業動向関連の情報抽出を行い、この手法の有効性を明らかにしてきた[1]。表層パターンに着目する手法は、読み手の深い知識がなくても文の表層構造から新しい情報が得られる書き方がされているという仮定に基づく。[2]では、専門用語や新語定義文にも限られた表層パターンがあるという仮説から、用語とその定義文の抽出を試みた。[3]では[2]を発展させ、新聞記事の表層情報から百科事典的用語集を自動構築するための分析を行った。分析は、「αはβ」型文に着目し、物事の性質を表す文を幅広く取る方針で行った。その結果、「αはβ」型文のβ中に「利用」と「される」または「できる」が出現する場合は<用途>を示す文、「購入」「合意」「進出」「派遣」「着手」などが出現する場合は<組織や人物の活動>を示す文が多いことがわかった。これらの中で最も多くの割合を占めたのは、<組織や人物の活動>を表す文だった。

そこで今回我々は、この<組織や人物の活動>の文に着目した。会社の年表や人物情報事典などの知識データベースを自動作成に使用する抽出規則を作成するために、述語を中心とした文末表現に着目して新聞記事テキストの分析を行った。日本語の文末には、述語動詞をはじめ、テンス、アスペクト、ヴォイスなど、文全体の意味にとって重要な要素が現れる。本稿では、新聞記事を用いて行ったテキスト分析の結果について報告する。

2 主な論点

[4]において我々は、文末表現を分析し、構造化テキスト検索システムの事象判定の方法を提案した。[4]では、記事1文目の文末表現に事象に特徴的な「行為語」が現れやすいことに着目し、「組織合併」「製品販売」について書かれた記事を分析している。そして、文末の述語表現は、第1層の行為語、第2層の開始語、第3層の発表語に分けて捉えられ、事象は第1層の行為語として表現されることを明らかにした。

本稿では、[4]で明らかになった(a)文末の述語表現が階層構造を持っており、(b)事象は行為語として表現されることを前提とする。その上で、

(1) 文末表現の階層構造：「組織合併」「製品販売」の事象では、文末述語の第1層に行為語があり、第2層に開始語があり、第3層に発表語があると捉えられた。他の事象では、この階層構造はどのように現れるのか。

(2) 知識データベースの作成方法：会社の年表や人物情報事典などの知識データベースに載せる価値のある情報を持った記事を集めるにはどうすればよいか。

の2つを主な論点とする。

3 調査

3.1 調査対象

分析には、1997年版毎日新聞の偶数月19日の記事を使用した。企業動向や人物の活動に関する記事が多そうな経済面記事217記事の1文目を調査した。半定型文の性質を持つ新聞記事では、1文目に記事全体の要約が述べられることが多い。そこで今回は、記事1文目のみを調査対象とした。

3.2 調査方法

以下の方法で調査を行う。

(1)記事の中から企業動向や人物の活動に関するものを選ぶ。

> フィルム要らずで手軽に写真が撮れ、パソコンで簡単にあいさつ状などに加工できるデジタルカメラがブームになっている。　（毎日新聞，1997年2月19日）

などはこの段階で排除する。

(2)行為主体、行為語、行為語に後続する要素にマークをつける。

> １８日付の<発表主体>米紙ワシントン・ポスト早版</発表主体>は、<行為主体>クリントン米大統領</行為主体>が国務次官への転出が予定されているスチュワート・アイゼンスタット商務次官（国際貿易担当）の後任として、デビッド・アーロン経済協力開発機構（ＯＥＣＤ）大使（５８）を<行為語>指名する</行為語>見通しだと<後続する述語>報じた</後続する述語>。

（毎日新聞，1997年4月19日）

(3)行為主体を、<組織>、<人物>などに分類する。その上で、行為主体の分類によって、共に出現する行為語や行為語に後接する要素に傾向があるかを見る。

(4)文末述語の階層構造には、行為語や行為主体ごとに接続の仕方や種類に傾向があるかを見る。

3.3 調査結果

分析対象とした217記事のうち、企業動向や人物の活動に関するものは141記事だった。この141記事の行為主体、行為語、行為語に後続する要素にマークをつけた。

[4]では「組織合併」「製品販売」の事象に限定した分析で、文末述語の第1層に行為語があり、第2層に開始語があり、第3層に発表語があることを明らかにした。本稿では、事象を限定せずに経済面記事の文末述語を調査した。その結果、行為語に後接する要素は、<事態の進行を表す動詞>、<事態の確実性を表す名詞>、<発表語>に分類できた。<事態の進行を表す動詞>は[4]の「開始語」に相当する。<事態の進行を表す動詞>には「開始する」「始めた」などの開始を表す語以外に、「新たな段階に入った」「行う」などが含まれる。<事態の確実性を表す名詞><発表語>は、それぞれ[4]の「体言化表現」「発表語」に相当する。[4]でも「発表語の述語表現形成パターン」として3つのパターンを挙げている通り、各要素の接続の仕方には幾通りかのパターンがある。[表1]に接続のパターンと該当数、[表2]に例を示す。

4 考察

(1) 文末表現の階層構造

「<行為語><事態の進行を表す動詞><事態の確実性を表す名詞><発表語>」の接続の仕方には、(a)6通りのパターンがあり([表1])、(b)行為語、事態の進行を表す動詞の時制によって変化する。

(a)の6通りのパターンは、[表1]に挙げたA型からF型を指す。最も多くの割合を占めたのは、F型の<行為語>のみが現れるもの(45.4%)、次に多かったのは<行為語>に<発表語>が接続したE型(24.8%)だった。

(b)の動詞の時制によって変化するものは、事態の確実性を表す名詞である。「こと」以外の「方針」「意向」「見通し」「考え」は動詞現在形、未来形(ル形)にのみ接続する。タ形に接続する「辞任した意向を表明した」という文は見られない。「こと」以外の「方針」「意向」「見通し」「考え」が、事柄が未確定であることを表すので、事態の完了を表す動詞タ形と共起すると不自然になる。一方、「こと」は事柄の確定性に対して中立的な名詞なので、動詞タ形ともル形とも共起できるのだろう。このことから、[表1]のA型とC型の文については、事態の確実性を情報として付与できる([図1])。

(2) 知識データベースの作成方法

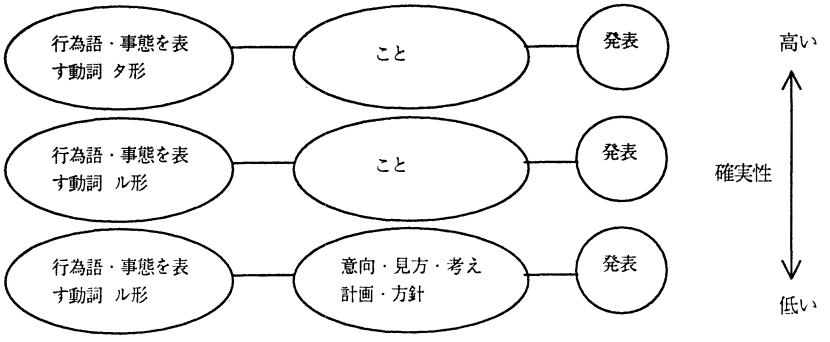
本稿では、事象は行為語として表現されることを前提としている。そこで、会社の年表や人物情報事典などの知識データベースに載せる事象を集めるには、行為語を集めればよい。では、行為語を集めるにはどのようにすればよいか。単純に文末表現だけを集めてくる方法では、

> お盆が終わって、大人たちは職場に復帰した。子どもたちはまだ夏休みの真っ最中だが、休みは残り少なくなった。遊んでばかりもいられない。
>
> （毎日新聞，1997年8月19日）

のような情報が取れることになる。行為主体が組織や人物になっているものに限定し、文末表現だけを集めたとしても、

> １８日夕の新進党両院議員総会は、結果的に小沢一郎党首が党首続投の意思を伝えるだけで議員たちの声を一切聞かず、３０分足らずで一方的に打ち切られた。
>
> （毎日新聞，1997年2月19日）

のような情報が取れる。この種の記事は、時間の経過とともに情報の価値が確実に薄れてくる。知識データベースに載せるべき情報ではなく、また「打ち切られ



[図1 ： 事柄の確実性]

［表１ ： 文末述語の接続のパターンと該当数］　　○:要素あり ／ 空欄:要素無し

型	事態の進行を表す動詞	事態の確実性を表す名詞	発表語	集計
A	○	○	○	4
B	○		○	6
C		○	○	20
D	○			12
E			○	35
F				64

［表２： 文末述語の接続のパターンと例］

事態の進行を表す動詞（有）－事態の確実性を表す名詞（有）－発表語（有）

行為主体	行為語	事態の進行を表す動詞	事態の確実性を表す名詞	発表語
組織	再編(体言)	を実施する	こと	を明らかにした
組織	提携(体言)	を行う	こと	明らかにした
人物	整備(体言)	を行う	方針	を示した

事態の進行を表す動詞（有）－事態の確実性を表す名詞（無）－発表語（有）

行為主体	行為語	事態の進行を表す動詞	事態の確実性を表す名詞	発表語
組織	サービス(体言)	を開始する		と発表した
組織	イベント(体言)	を始める		と発表した
組織	手続き(体言)	の開始		を決定

事態の進行を表す動詞（無）－事態の確実性を表す名詞（有）－発表語（有）

行為主体	行為語	事態の進行を表す動詞	事態の確実性を表す名詞	発表語
組織	指定する(用言)		こと	を承認した
人物	就航させる(用言)		計画	を発表した
人物	廃止する(用言)		こと	を明らかにした
組織	交わしていた(用言)		こと	を明らかにした
人物	歓迎(体言)		の意向	を表明した
人物	辞任する(用言)		意向	を表明した
人物	設ける(用言)		方針	を表明した
人物	歓迎する(用言)		考え	を示した
人物	必要になる(用言)		見方	を示した

事態の進行を表す動詞（有）－事態の確実性を表す名詞（無）－発表語（無）

行為主体	行為語	事態の進行を表す動詞	事態の確実性を表す名詞	発表語
組織	サービス(体言)	を始める		
組織	発売(体言)	を始めた		
組織	事業(体言)	を始める		
組織	合併交渉(体言)	(が) 新段階に入った		
組織	創設(体言)	で大筋合意した		

事態の進行を表す動詞（無）－事態の確実性を表す名詞（無）－発表語（有）

行為主体	行為語	事態の進行を表す動詞	事態の確実性を表す名詞	発表語
組織	決定する(用言)			と発表した
組織	建設する(用言)			と発表した
組織	発売する(用言)			と発表した
組織	合併する(用言)			と発表した
組織	造る(用言)			と発表した
組織	設立した(用言)			と発表した
組織	決算(体言)			を発表した
組織	人事(体言)			を発表した
組織	リポート(体言)			を発表した
組織	計画(体言)			を発表した
組織	見通し(体言)			を発表した
組織	コメント(体言)			を発表した
組織	結果(体言)			を発表した
人物	不満(体言)			を示した
人物	考え(体言)			を示した
人物	姿勢(体言)			を示した

た」も行為語か否かの判断が難しい。確実に行為語である表現を集める必要がある。

本稿では、[表 1]に示した文末述語の接続パターンに着目して、行為語を抽出する方法を提案する。「と発表した」「を明らかにした」などの発表語は、半定型文と言われる新聞記事に多く見られる表現である。新聞記事の分析作業中に、文末に発表語が現れる文は行為語と共出している傾向が強いことがわかってきた。そこで、

① 経済面に限定せず、毎日新聞の全記事から頻度の高い文末表現を抜き出した([表 3])。その中から、発表語の候補を挙げる。今回は、「発表した」「明らかにした」「示した」を候補とする。

② 文末に発表語がある記事を選び出し、そこから<行為語>を抽出する。[表 1]に示した文末述語の接続パターンのうち、A,B,C,E 型が文末に発表語が出現するパターンである。A 型なら発表語の 3 階層前、B,C 型なら 2 階層前、E 型なら 1 階層前に行為語がある。「事態の進行を表す動詞」「事態の確実性を表す名詞」は種類が限られているので、登録しておける。これにより、形態素解析を用いない字面のみのパターンマッチングで行為語を抽出できる。

③ ②の作業によって、行為語(事象)の候補が挙がる。

④ ③であげた候補を元に、

(a)会社の年表や人物事典に載せる事象を選ぶ

(b)抽出規則を作成する

という方法を提案する。

5 おわりに

以上、会社の年表や人物情報事典などの知識データベース自動作成用の抽出規則を作成するために、述語を中心とした文末表現に着目して新聞記事テキストの分析を行った。分析の主な論点は、(1)文末表現の階層構造、(2)知識データベースの作成方法、の 2 点とした。実作業および抽出規則作成は、今後の課題としていきたい。

参考文献

[1] 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, 情処研報, NL124-23, pp.95-102 (1998).

[2] 西野文人, 橋本三奈子, 落谷亮：テキストからの用語とその定義文の抽出, 言語処理学会 第 5 回年次大会 発表論文集, pp.124-127 (1999).

[3] 木田敦子, 乾裕子, 落谷亮, 西野文人：新聞記事からの用語集作成のためのテキスト分析, 情処研報, NL134-12, pp.85-92 (1999).

[4] 桑畑和佳子, 橋本三奈子, 木田敦子, 落谷亮, 西野文人：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, 言語処理学会 第 4 回年次大会 発表論文集, pp.634-637 (1998).

参照資料

[5] CD-毎日新聞データ集('91～'97), 毎日新聞社.

[表 3： 毎日新聞紙面全体での 20 位までの文末表現頻度]

順位	%	累積%	頻度	文末表現
1	5.906	5.906	37910	)
2	3.751	9.657	24074	発表した。
3	2.252	11.909	14455	明らかにした。
4	1.976	13.884	12681	決めた。
5	1.640	15.525	10529	。
6	1.474	16.998	9460	なった。
7	1.062	18.060	6816	分かった。
8	1.012	19.072	6493	」。
9	0.960	20.032	6163	」
10	0.859	20.891	5511	――。
11	0.737	21.628	4731	い。
12	0.693	22.321	4448	)。
13	0.678	22.998	4350	行った。
14	0.642	23.640	4118	行われた。
15	0.630	24.270	4045	逮捕した。
16	0.628	24.898	4032	まとめた。
17	0.618	25.516	3965	開かれた。
18	0.599	26.115	3845	示した。
19	0.560	26.675	3594	始まった。
20	0.547	27.222	3513	た。