情報抽出のための新聞記事テキスト分析[1]

木田敦子、乾裕子 *1　桑畑和佳子、橋本三奈子 *2　落谷亮、西野文人 *3
*1 計量計画研究所　　*2 富士通　　*3 富士通研究所

1 はじめに

近年、新聞記事テキストから情報抽出を行う研究がさかんに行われている。これらの研究では、表層情報を利用した抽出規則データによって情報抽出を行っている。

我々は従来の情報抽出に加え、新聞記事テキストから事象構造を抽出する方法について研究を進めている。これは、シナリオテンプレートの考え方に基づく。シナリオテンプレートとは、取り出すべき情報をあらかじめ決定しておき、それに基づいて抽出する手法である。これにより、テキストに現れた表層情報の抽出ではなく、事象構造の抽出を実現する。この方法は、新聞が半定型的テキストであることを利用したものである。企業動向などに関する新聞記事は直感的にも半定型的であると考えられるが、その構造の詳細については十分検討されていなかった。本稿では、新聞記事の半定型的構造を明らかにするために、企業動向に関する記事の一文目と二文目以降の関係を調査する。あわせてシナリオテンプレートによる情報抽出への提言を行う。

2 事象と文構造の対応

我々は半定型文書の性質を持つ新聞記事から事象構造を取り出す情報抽出実験を行っている。抽出実験に先立つ予備調査を経て、事象構造と文構造には相関関係が見られることが明らかになってきた。以下に挙げるのは、記事一文目の事象構造と文構造の関係を調査した結果である。組織合併情報、製品販売情報についての結果を例に示す。組織合併、製品販売ともに、日経新聞の200記事を調査対象とした。

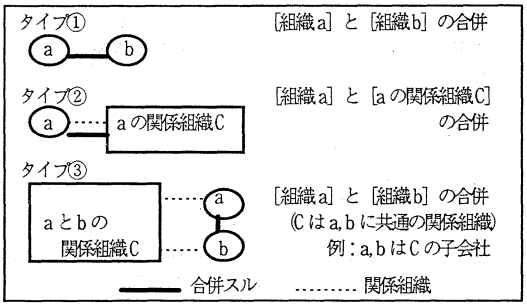

図1 組織合併の事象構造

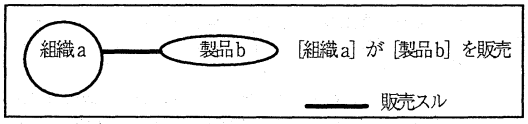

図2 製品販売の事象構造

表1 組織合併の文構造

a:[合併元以外の組織名] ハ/ガ [会社タイプ][数社] ヲ 合併
b:[合併元以外の組織名] ハ/ガ [会社タイプ][数社] ヲ <組織名> ニ 合併
c:[合併元以外の組織名] ハ/ガ <組織名> ヲ <組織名> ニ 合併
d:[合併元以外の組織名] ハ/ガ <組織名> ト <組織名> ヲ 合併
e:[数] [団体タイプ] ガ 合併
f:[団体タイプ] ハ/ガ 合併
g:[組織タイプ] [数] ハ/ガ 合併
h:<組織名> ハ/ガ [組織タイプ] [数社] ト 合併
i:<組織名> ハ/ガ [組織タイプ] [数社] ヲ 合併
j:<組織名>など [数社] ハ/ガ 合併スル
k:<組織名> ハ/ガ <組織名> ト 合併
l:<組織名> ト <組織名> ハ/ガ 合併
m:<組織名> ハ/ガ <組織名> ヲ 合併
n:<組織名> ハ/ガ <組織名> ニ 合併 (サレタ)
o:<組織名列挙> ハ/ガ 合併
* 組織タイプ=子会社；関連会社；グループ会社　など
* 団体タイプ=農協；漁業組合；農業組合　など

表2 製品販売の文構造

a:<組織名> ハ/ガ <製品名> ヲ 販売
a':<組織名> ハ/ガ <製品名> ノ 販売
b:<組織名> ハ/ガ <製品名> ヲ [動詞] 販売
c:<組織名> ハ/ガ [共同組織] ト <製品名> ヲ 販売
c':<組織名> ハ/ガ [共同組織] ト <製品名> ノ 販売
d:<組織名> ハ/ガ <製品名> ヲ [共同組織] ト [動詞] 販売
d':<組織名> ハ/ガ [共同組織] ト <製品名> ヲ [動詞] 販売

製品販売の事象構造が一種類であるのに対し、組織合併の事象構造は三種類に分かれる。文構造については、製品販売が七種類であるのに対し、組織合併は十六種類に分かれる。このことから、事象構造の多様化とともに文構造も多様化することがわかる。また、組織合併の例では、事象タイプ②に文構造m型が多く、事象タイプ③に文構造a型が多い(図3)。このように、事象構造と文構造には相関関係が認められる。この予備調査の結果は、比較的単純なパターンの照合で、文構造から事象構造を取り出せる可能性を示している。そこで我々は、事象－表現マッピング規則、語彙リストを用いてトップダウン式にパターン解析を行うシステムを作成し、抽出実験を行った（システムの詳細については[2]を参照のこと）。システムの構築とともにマッピング規則や語彙リストを記述し、抽出規則データを作成した。この抽出規則データは、一文目の情報を中心に作成している。

この過程で、我々は次の仮説を立てた。新聞記事の場合、一文目に記事全体の要約が半定型的な形で記述され、二文目以降に一文目の補足となる情報が記述されている。この仮説

1 本研究は、平成9年度富士通株式会社の委託調査研究によるものである。

を検証するために、1)二文目以降が一文目に対してどのような位置づけにあるのか、2)二文目以降がどのような情報を持っているのか、3)二文目以降の情報は記事内容によって変わるのかを明らかにする。まず、事象構造を把握するのに一文目の情報だけで十分なのかという疑問を解決する。そのために、組織合併情報、製品販売情報について、新聞記事の一文目に現れる情報、二文目以降に現れる情報の調査を行った。

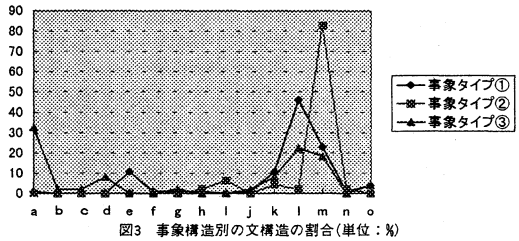

図3 事象構造別の文構造の割合(単位:%)

3 新聞記事テキストの調査

3.1 調査対象

組織合併情報、製品販売情報を示す特徴語(例:「販売」「発売」「合併」)を一文目文末の述語に持つ新聞記事について調査した。日経新聞から、組織合併情報、製品販売情報に関する200記事をそれぞれ取り出し、調査対象とする。200記事の内訳は、95年分から100記事、96年分から100記事とする。

3.2 調査方法

以下の項目の出現位置(新聞記事の何文目に出現したか)と出現の仕方(全項目出現したか、一部の項目のみ出現したか)を調べる。

組織合併情報
- ・合併元組織名
- ・新組織名

製品販売情報
- ・販売元組織名
- ・販売製品名
- ・価格

3.3 調査結果

調査結果を下記に示す。出現した位置と出現の仕方によって以下の基準(表3)で記号を入れ(表4)、これをグラフ化した(図4)。

調査結果から、記事一文目に出現する項目と二文目以降に出現する項目の相違が明らかである。合併元組織名、販売元組織名、販売製品名は一文目に出現する傾向にある。これに対し、新組織名や価格は二文目以降に出現するか、まったく出現しない傾向にある。

表3 出現傾向の基準

出現位置 / 出現の仕方	一文目	二文目以降	一文目及び二文目以降
全項目出現	a1	a2	a3
一部の項目のみ出現	b1	b2	b3
まったく出現せず	c		

4 シナリオテンプレート

4.1 シナリオテンプレートとは

シナリオテンプレートとは、最初に取り出すべき情報を決定しておき、そこで必要とされた情報を抽出する規則データである。我々が最初の段階で作成していた一文目中心の抽出規則データは、文を解析して文中の要素に属性タグを付与するものであった。組織合併を表す「A社とB社が合併する」という文に対し、「A社」「B社」に<合併元組織名>、「合併する」に<組織合併行為>という属性タグを付与する。これに対しシナリオテンプレートは、あらかじめ埋めるべき属性タグが決まっている。例えば組織合併を表す文であれば、<合併元組織情報>、<組織合併行為>、<新組織名>という空きスロットが用意されており、文を解析していく過程で順次スロットが埋められていく。

シナリオテンプレートの手法では、新組織名のように二文目以降に出現する情報も抽出することができる。そこで、我々は次のような方法で抽出実験を進める。

(1)一文目を解析し、その記事が何の事象について書かれているかを明らかにする(これに関しては、[3]参照のこと)。
(2)埋めるべきスロットが決まる。
(3)一文目の文全体を解析しても埋められなかったスロットは、二文目以降のキーワードや文構造などを手がかりにして、必要な情報だけを探して埋めていく。

本節では、抽出実験全体の中から、(3)の二文目以降の情報抽出実験とその報告を行う。

4.2 抽出規則データの作成方法と作成対象

一文目に対して行ったような全文解析をせずに情報を抽出するには、以下のような方法が考えられる。

(1)パターンマッチングで二文目以降から情報抽出する。
(2)一文目を解析して得られた情報をもと、に二文目以降から情報抽出する。

表4 出現傾向

調査項目		a1	a2	a3	b1	b2	b3	c
組織合併情報	合併元組織名	159	11	0	7	1	0	22
	新組織名	0	61	0	0	0	0	139
製品販売情報	販売元組織名	195	0	0	0	0	0	5
	販売製品名	188	1	6	3	0	0	2
	価格	3	45	0	0	0	0	152

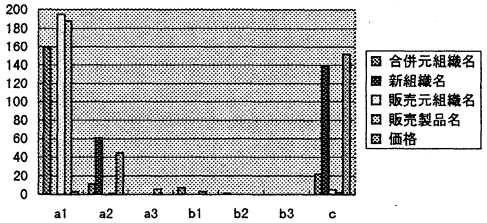

図4 項目別の出現位置(単位:該当記事数)

我々は、(1)のパターンマッチングによる二文目以降からの抽出方法で実験を行った。対象とした抽出項目は、二文目以降に出現する新組織名、価格、合併元組織名である。これらの抽出項目が二文目以降に出現していた記事(3.3 調査結果 表4 a2)を分析し、抽出規則データを作成した。分析に使用した記事数は以下の通りである。

<新組織名>　：61 記事
<合併元組織名>　：11 記事
<価格>　：45 記事

分析結果をもとに作成した抽出規則データで抽出した結果をさらに分析するという再帰的な方法で、抽出規則データに修正を加え精度向上を図った。抽出規則データは、新組織名、合併元組織名、価格の各項目を抽出するために各々用意する。各抽出規則データ作成に要した時間は、それぞれ3時間である。3 時間に達したときに修正が十分でなかった場合も、その段階で作業を打ち切った。

4.3 分析

◇ 新組織名

キーワード

① 新名称は，新社名は，新会社名は，新会社の名前は，新会社の社名は，新会社の名称は，合併後の社名は，合併後の名称は，新会社は，存続会社は
② 新会社である，新会社の
③ 新.+?名は
④ 新設{さ|し|す}，発足{さ|し|す}
⑤ 社名を～に

組織合併事象を表す記事の二文目以降では、①のキーワード群と「と」「で」「に」「を」「が」「、」「。」の間にある任意文字列は、新組織名である可能性が高い。また、直後に「が」「を」があり、それに連続して④のキーワード群がある任意文字列も新組織名である可能性が高い。このように分析を進めていき、大きく分けて五種類のキーワード群を用意した。これをルール化し、抽出規則データを作成した。

◇ 合併元組織名

キーワード

① 合併するのは，対象は
② ( ) 、，( ) と
③ {合併|吸収|吸収合併}{さ|し|す}
④ {合併|対等合併}{し|す}

組織合併事象を表す記事の二文目以降では、①のキーワード群と「で」「。」の間の任意文字列が②のキーワードを含んでいれば、合併元組織名である可能性が高い。また、「αがβを」の後に③のキーワードがある場合、「αがβと」の後に④のキーワードがある場合、任意の文字列α、βは合併元組織名である可能性が高い。このように分析を進めていき、大きく分けて四種類のキーワード群を用意した。これをルール化し、抽出規則データを作成した。

◇ 価格

キーワード

① 価格は
② 円，ドル
③ 台前半，台後半，台，以上，以下
④ [〇一二三四五六七八九十百千万億兆両０１２２３４５６７８９]の任意回繰り返し

製品販売事象を表す記事の二文目以降から価格を抽出する場合、上記の①から④に挙げた語がキーワードになる。だが価格の場合、キーワード群を用意して、それをルール化するだけではうまくいかない場合が多い。以下の記事のように、価格が場合分けで出てくることがある。

> 清酒「福正宗」を製造・販売している福光屋（金沢市、福光松太郎社長）は、しぼりたての本醸造・生原酒をそのまま直詰めした「福正宗しぼりたて生・原酒」＝写真＝と吟醸新酒「一九九五」を発売した。「しぼりたて生・原酒」は火入れや調合など一切の手を加えていないため、酒蔵から出したばかりの味わいが楽しめるという。価格は通年発売のアルミ缶が二百ミリリットル入りで三百円、冬季限定販売のガラス瓶が三百ミリリットル入りで四百五十円となっている。アルミ缶が年間十万本、ガラス瓶が一万本の販売を見込んでいる。
> [日経新聞CD-ROM版 95年版]

そこで、まず記事二文目以降から「価格」という文字列を含む文を取り出す。この文がなければ「円」「ドル」を含む文を取り出す。そして、取り出した文をパターンマッチングで解析するという方法をルール化し、抽出規則データを作成した。だが、以下の記事のように価格が二文にわたって出現する場合は抽出漏れが出るなど、問題は残っている。

> 酒造メーカーの六歌仙（東根市、ＴＥＬ０２３７・４２・２７７７）は一月末までしぼりたての吟醸酒「しぼりたて吟醸」を宅配販売する。原料米に美山錦と雪化粧を使い醸造した吟醸酒をしぼりたてで生のまま瓶詰めした。火入れせず、やや炭酸ガスを含んださわやかさが特長。送料、消費税込み五百ミリリットル瓶二本セット三千円、三本セット四千円。吟醸酒二本とそば五人前のセットは五千円。　[日経新聞CD-ROM版 95年版]

4.4 実験結果

作成した抽出規則データを使い、抽出実験を行った。実験結果を以下に示す(表5)。

表5　実験結果

	抽出成功	一部抽出失敗	抽出失敗	合計
新組織名	59 (97%)	0 (0%)	2 (3%)	61
合併元組織名	9 (82%)	0 (0%)	2 (18%)	11
価格	34 (76%)	7 (15%)	4 (9%)	45

◇ 新組織名の抽出に失敗したもの

・新組織名が複数出現するもの

> ～。これにより「郡山市」「すかがわ岩瀬」「あいづ」「会津みなみ」「そうま」の五農協が発足する。

[日経新聞CD-ROM版96年版]

・キーワードが重なるもの

——。小木曽工務所が存続会社で、合併比率は小木曽株一株に対して、松尾建一株、熊谷産〇・二株。入札制度の見直しなど経営環境の変化に対応し、若年労働者の確保と民間建築工事の受注獲得を目指す。新会社は官公需依存の体質から脱皮し、二〇〇〇年初頭には株式を公開したい考えだ。　[日経新聞CD-ROM版95年版]

◇ 合併元組織名の抽出に失敗したもの

・キーワードがないもの

◇ 価格の抽出に失敗したもの

・修飾句中に「〇〇円(ドル)」が含まれるもの

——、通常料金より千円安い三千三百円で購入できる。
[日経新聞CD-ROM版95年版]

・価格が範囲指定で示されるもの

——。価格は一メートル当たり百五十一二百円。
[日経新聞CD-ROM版96年版]

5 考察

記事一文目だけに着目すると、製品販売情報と組織合併情報では、組織合併情報の方が複雑な事象構造と文構造を持っている。一方、二文目以降では、組織合併情報よりも製品販売情報の方が多様な情報を持ち、文構造も複雑である。

このように一口に企業情報と言っても、その情報の種類により、新聞での記事の書き方が異なることが明らかになった。これは、抽出する情報の種類により、抽出方法を変えるべきことを意味する。

以下の記事では、一文目から販売製品は「スーツ地」であると判断できる。だが、以降の文で出てくる価格は、製品であるスーツ地の価格ではなく、「仕立て上がり価格」である。

御幸毛織は、絣（かすり）柄のスーツ地を九六年春夏物として販売する。ウール六五%、モヘア三五%を素材にした。これらの繊維による絣柄の生地は珍しいという。仕立て上がり価格はスーツで約二十九万円、ジャケットで約二十万円。　[日経新聞CD-ROM版96年版]

また、次の記事では、一文目に出ている販売製品は「婦人服の新ブランド『マーク・ジェイコブス・ルック』」である。だが、以降の文に「コート」「ジャケット」「スカート」と製品の詳細情報が出ており、それぞれの価格が示されている。

レナウンルックは人気服飾デザイナー、マーク・ジェイコブス氏と提携、今秋から婦人服の新ブランド「マーク・ジェイコブス・ルック」を販売する。シンプルで繊細なデザインが特徴。中心価格はコート五万四千円、ジャケット四万二千円、スカート一万五千円と普及価格帯に抑えた。初年度の売り上げ目標は小売りベースで二十五億円。　[日経新聞CD-ROM版96年版]

価格が量の単位とともに示される記事もある。「二百ミリリットル入りで三百円」、「四百五十四グラム瓶入りで、希望小売価格は各五百五十円」などが例として挙げられる。これらは、二文目以降で新たな情報が加えられるケースである。

これに対して、組織合併情報の多くは、二文目以降が一文目に現れた情報の補足になっている。二文目以降に新組織名や合併元組織名が現れる記事もあるが、出現する情報に多様性はなく、現れ方も定型的である。そのため、二文目以降からの情報抽出が、製品販売情報に比べると容易である。

また、シナリオテンプレート作成にあたっても、組織合併情報の方は出現する情報に多様性がないため、埋めるべき項目を決定しやすい。また、一文目の情報を使って二文目以降から情報抽出を行うことも可能であると考える。以下の記事では合併元組織名が二文目以降に現れている。この記事には「合併するのは」や組織補足情報を表す「（）」のようなキーワードがないが、一文目の「六信用金庫」を使うことによって、キーワード「信用金庫」を含む文字列六種類という探し方ができる。

東京、青森、広島の六信用金庫が相次いで合併する。東京が朝日信用金庫と浅草信用金庫、青森が北奥羽信用金庫と青森信用金庫、広島が鞆（とも）信用金庫と福山信用金庫の合計三組。いずれも八—十月にかけて対等合併する。一方、福岡では行橋信用金庫が北九州八幡信用金庫に事業譲渡し、解散する。信用組合などでは経営破たんが相次いでいるが、八信金は合併・事業譲渡という手段で経営基盤を強化することで、生き残りを図る方針だ。
[日経新聞CD-ROM版96年版]

一方、製品販売情報の方は、一文目の文構造が複雑でない分、二文目以降の抽出に使える情報が一文目に現れていないのである。

6 おわりに

以上、企業動向に関する記事を対象に、一文目と二文目以降の関係を調査し、新聞記事の半定型的構造について述べた。そして、その半定型的構造をふまえた上で、シナリオテンプレートによる情報抽出への提言を行った。今回の調査で得られた知見に基づくシナリオテンプレートの精度向上を今後の課題とする。

引用文献

[1] 日経全文記事データベース 日本経済新聞CD-ROM版 1995年版

[2] 日経全文記事データベース 日本経済新聞CD-ROM版 1996年版

参考文献

[1] 井手裕二, 藤吉誠, 永井秀利, 中村貞吾, 野村浩郷. 構造化テンプレートを用いた新聞記事からの製品情報抽出. 情報処理学会研究会報告, NL118-2, 1997.

[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. 情報処理学会研究会報告, 1998-3.

[3] 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, 1998.