修　士　論　文

相対表現に基づく
動向情報抽出機構に関する研究

平成20年度修了
三重大学大学院　工学研究科
博士前期課程　情報工学専攻

上西 康広

要旨

WWW に代表されるように, 情報洪水の中から有用な情報のみを効率良く選択して利用することは現代社会において非常に重要であるが, 同時に非常に難しい問題でもある. こうした「玉石混合」から「玉」を取り出すための技術として, 「動向情報の要約・可視化」が注目されている . 「動向情報」とは, 幾つかの統計量の時系列データを基として, その変化を通時的にとらえつつ, それらを単に羅列するのではなく, 総合的にまとめ上げることで得られるもので, ある商品の価格や売上の状況, ある会社の業績状況, 内閣や政党の支持状況等が例としてあげられる. 動向情報は, 時系列データや地理的な情報を多く含み, さらに, それらの情報に対する解釈, 原因や予測などの情報を含むという性質を持つ. したがって, グラフや図, 地図などを利用してまとめて視覚化した方が直感的に理解し易い. 「動向情報の要約・可視化」技術はテキスト情報から動向情報を自動抽出し, ユーザの関心に応じて最適な視覚情報として再構成することにより, このようなニーズに対応できる技術である.

動向情報の一部として現れる統計量や日付表現には, 「前年比 10%増」のような数値の相対的な差異や, 数値の変動を示す相対表現が存在する. 相対表現には, 他の情報と対応付けることができれば, テキスト中に明示されていない情報を推論することができるという特徴がある. これまでに, 今岡ら (今岡ら 2006) は相対表現を利用した動向情報抽出の有効性を確認しているが, しかしながら, 彼らの手法は, 抽出した情報の選択については, 議論の対象外としていた. 動向情報を適切に抽出するためには, 抽出した情報の選択は不可欠である. なぜなら, 抽出した情報を選択しなければ, 抽出の対象となる情報以外の情報も抽出してしまう恐れが生じるからである. 例えば, 上で挙げた「ビールの出荷量」に関する動向情報を抽出する場合を考えると, テキスト中に存在する様々な種類の動向情報の中からビールの出荷量に関する情報を選択し, 抽出しなければならない. このような問題を解決するため, 2 種類の 4 つ組選択手法を提案する. 第一の 4 つ組選択手法は, 電子化辞書と人手で作成した同義語辞書および関連語辞書により specifier を拡張するという特徴がある. 第二の 4 つ組選択手法は, 入力となる統計量名 (クエリ) に対して, 全ての形態素の組合せを作成し, さらに同義語辞書を用いてクエリのバリエーションを増加させるという特徴がある.

本研究では, 提案する 2 つの 4 つ組選択手法をそれぞれ組み込んだ相対表現に基づく動

向情報抽出手法を提案する. 新聞記事を用いて評価実験したところ, 第一の 4 つ組選択手法を組み込んだ手法は適合率 0.677, 再現率 0.367, F 値 0.476(マイクロ平均) という性能が得られた. 第二の 4 つ組選択手法を組み込んだ手法は関連性判定を表層的な包含関係とした場合, 適合率 0.615, 再現率 0.397, F 値 0.483(マイクロ平均) という性能が得られた. 関連性判定を完全一致とした場合, 適合率 0.848, 再現率 0.367, F 値 0.512(マイクロ平均) という性能が得られた. このことから, 提案する相対表現に基づいた動向情報抽出の有効性が明らかとなった.

目 次

第1章

序論

1.1 研究の背景と目的

IT が促進したことによって, 様々な電子化情報が増加し続けている. 情報洪水の中から有用な情報のみを効率良く選択して利用することは現代社会において非常に重要であるが, 同時に非常に難しい問題でもある. こうした「玉石混合」から「玉」を取り出すためには, ユーザの関心に応じた柔軟な情報編纂技術 (加藤恒昭 松下光範 2006) が必要である. 情報編纂を指向する研究テーマとして, 「動向情報の要約・可視化 (KATO, MATSUSHITA, and KANDO 2005)」が注目されている. 「動向情報」とは, 幾つかの統計量の時系列データを基として, その変化を通時的にとらえつつ, それらを単に羅列するのではなく, 総合的にまとめ上げることで得られるもので, ある商品の価格や売上の状況, ある会社の業績状況, 内閣や政党の支持状況等が例としてあげられる. 動向情報は, 時系列データや地理的な情報を多く含み, さらに, それらの情報に対する解釈, 原因や予測などの情報を含むという性質を持つ. したがって, 文章で説明するよりも, グラフや図, 地図などを用いて様々な情報をまとめて視覚化した方が直感的に理解し易い. 動向情報を効率的に把握する支援技術として, 文書情報を解析して動向情報を自動抽出し, ユーザの関心に応じて最適な視覚情報として再構成する技術は非常に有効である. 文書中に記述された動向情報を抽出して可視化するためには, 以下のようなステージが必要である.

stage 1: 可視化の基本となる情報の抽出

stage 2: 関心や情報の種類に応じた可視化形式の選択

stage 3: 注釈などの文書情報を用いた可視化情報の補足

上にあげた 3 つのステージのうち, 最も基本的なものが stage 1:可視化の基本となる情報の抽出である. これらの基本的な情報を正確に抽出するためには, 文書中に記述された動向

情報, すなわち動向の種別を意味する表現や, それらに関連する数量表現や時間表現を見つけ出し, 相互に関連付けを的確に行う必要がある. 例えば, 「今年の月毎のビール出荷数量の推移」を処理しようとした場合, まず, 対象とする文書集合から, ビール出荷数量を意味する表現と出荷数量を示す数量表現と月を示す時間表現を抽出しなければならない. 次に, 月を示す時間表現とビール出荷数量を示す数量表現の対応関係を把握し, 月毎のビール出荷数量を12ヶ月分揃えることで, 目的とする動向情報が把握できる.

動向情報の一部として表れる統計量や日付表現には, 「前年比10%増」のような数値の相対的な差異や, 数値の変動を示すものがある (図1-1, 下線部). 難波ら (難波英嗣, 国政美伸, 福島志穂, 相沢輝昭, 奥村学 2005) は, これらの表現を**相対表現**と呼んでいる. 相対表現は, 抽出して他の情報と対応付けることができれば, テキスト中に明示されていない情報を推論することができるという特徴がある.

(1) 2007年のアサヒのビール出荷量は**前年比0.1%増**の1億8824万ケースとなった。

(2) 日経平均株価は**前日終値比218円33銭安**と続落し…

(3) **97年度末に比べ36万4000台減**の636万3000台となった.

図 1-1 相対表現の出現例

今岡ら (今岡裕貴, 桝井文人, 河合敦夫, 井須尚紀 2006) は, 新聞記事中の相対表現の出現の仕方に関する調査結果に基づいて抽出規則を構築した. 構築した抽出規則によって動向情報を抽出する実験を行い, 相対表現の有効性を明らかにしている. しかしながら, 彼らの手法は, 抽出した情報の選択については, 議論の対象外としていた. 動向情報を適切に抽出するためには, 抽出した情報の選択は不可欠である. なぜなら, 抽出した情報を選択しなければ, 抽出の対象となる情報以外の情報も抽出してしまう恐れが生じるからである. 例えば, 上で挙げた「ビールの出荷量」に関する動向情報を抽出する場合を考えると, テキスト中に存在する様々な種類の動向情報の中からビールの出荷量に関する情報を選択し, 抽出しなければならない.

このような問題を解決するため, 2種類の4つ組選択手法を提案する. 第一の手法は, 電子化辞書と人手で作成した同義語辞書および関連語辞書によりspecifierおよびheadwordを拡張するという特徴がある. 第二の手法は, 入力となる統計量名 (クエリ) に対して, 全ての形態素の組合せを作成し, さらに同義語辞書を用いてクエリのバリエーションを増加させるという特徴がある. 本論文では, これらの二種類の手法を組み込んだ相対表現に基づく動向情報抽出手法について述べる.

1.2 論文の構成

本論文は, 全8章で構成されている. 第1章の序論に続き, 第2章では, 動向情報抽出に関する研究が行われているMuSTプロジェクトについて説明する.

第3章では, 本研究で扱う動向情報の基本要素について説明する. また, 相対表現と基本要素との関連について説明し, 相対表現の種類について述べる.

第4章では, 相対表現を利用した動向情報抽出の基本的な処理の流れについて説明する. また, 4つ組選択の問題点について述べる.

第5章では, 第4章で述べる4つ組選択における問題点を解消するための第一の4つ組選択手法について述べる. 本手法は, 電子化辞書と人手によって作成した同義語辞書と関連語辞書を用いて4つ組の選択を行う.

第6章では, 第二の4つ組選択手法について述べる. 本手法は, 入力となる統計量名 (クエリ) に対して, 全ての形態素の組合せを作成する. さらに同義語辞書を用いてクエリのバリエーションを増やして4つ組の選択を行う.

第7章では, 二つの4つ組選択手法を第4章で述べる相対表現を利用した動向情報抽出に組み込んだ手法の有効性を検証するために評価実験を行い, その結果について考察する.

第8章では, 本研究に関する結論を述べる.

第2章

MuST

本章では, 動向情報抽出に関する研究が行われている MuST プロジェクトについて説明する.

2.1 MuST プロジェクト

動向情報に関する研究は, 2005 年に「MuST(A Workshop on Multimodel Summarization for Trend Information)」という動向情報の要約と可視化に関するワークショップがスタートし (KATO et al. 2005), そこで様々な研究が行われている.

MuST ワークショップは, 共通のデータを用いて, 緩い意味で共通の課題に取り組むことによる議論と研究の活性化, ツールやコーパス類の蓄積を目的とするもので, 2005 年にスタートした. MuST は NII NTCIR-7 タスクとして位置づけられている.

MuST では, 主にコーパスからどのように動向情報を抽出するかという研究と, 動向情報をどのように可視化するかという研究が行われている.

動向情報抽出に関する研究として, Yoshida et al(Yoshida, Sugiura, Hirokawa, Yamada, Masuda, and Nakagawa 2008) や Nanba et al(NANBA, OKUDA, and OKUMURA 2007) の研究が挙げられる. Yoshida et al は, 新聞記事に対して機械学習を利用して統計量名や数値情報の抽出する研究を行っている. Nanba et al は, 対象を新聞記事から blog に拡張して, 動向情報抽出の研究を行っている.

可視化に関する研究として, Takama et al(Takama 2007) や松下ら (松下光範 加藤恒昭 2007) の研究が挙げられる. 高間らは, 地震情報を日本地図や折れ線グラフなどの複数の可視化表現を用いてユーザに提示する手法を研究している. 松下らは, 統計 DB 等から得られる時系列数値情報と、それに関連する内容の一連のテキストを関連付けて視覚化し、ユーザの探索的データ分析を支援する可視化インタフェースについて研究している.

2.2 MuSTコーパス

MuSTでは,動向情報研究用のタグ付きコーパス(MuSTコーパス)を公開している.MuSTコーパスは,毎日新聞電子化版(1998年,1999年)の2年分から,パソコンの出荷状況や内閣支持率など,あらかじめ設定された27トピックに関する記事を抜き出し,XMLで定義された動向情報を示す13種類の要素に対して人手でタグを付与したテキストコーパスである.例として,トピック「パソコン」に関するMuSTコーパス記事を図2-1に示す.図中,"<TEXT> ··· </TEXT>" に囲まれた箇所が記事である.記事中では,<name>タグが付与される部分は統計量名,<date>タグが付与される部分は日付,<val>タグが付与される部分は統計量,<rel>タグが付与される部分は,統計量の値の差や比,順位等の相対値に相当する.

```
<?xml version="1.0" encoding="Shift_JIS"?>
<DOC>
<DOCNO>010803062</DOCNO>
<SECTION> 経済 </SECTION>
<AE> 有 </AE>
<WORDS>623</WORDS>
<HEADLINE><unit stat="国内出荷台数"><date gra="四半期" abs="200106"> 4～6月期 </date> の <name part="foot"> パソコン出荷 </name>、伸び率=急ブレーキ－－ <name part="head">「台数」</name><rel type="other"> 頭打ち </rel></unit></HEADL>
<TEXT>
  <unit stat="国内出荷台数"> 電子情報技術産業協会（ＪＥＩＴＡ）が2日発表した <date gra="四半期" abs="200106"> 今年度第1四半期（4～6月）</date> のパソコン出荷実績によると、<name> 国内向け出荷台数 </name> は <date gra="四半期" abs="200006"> 前年同期 </date> 比 <rel type="prop"> ２％ </rel> 増の <val> ２７９万３０００台 </val> と、これまでの <rel type="prop"> ２ケタ </rel> 増から一転、<date gra="四半期" abs="199809"> ９８年７～９月 </date> 以来 <dur gra="年"> 約３年 </dur> ぶりに <rel type="prop"> １ケタ </rel> 増にとどまった </unit>。一般消費者向けが１０％も減少したことが響いた。
</TEXT>
</DOC>
```

図 2-1 MuSTコーパスの例

第3章

動向情報と相対表現

本章では, 本研究で扱う動向情報の基本要素について説明する. さらに, 相対表現と基本要素との関連について説明し, 相対表現の種類について述べる.

3.1 基本要素と相対表現

文書中から動向情報を適切に抽出するためには, その基本となる要素を定義しておく必要がある. 本研究では, MuST コーパスに従って基本要素を {**name(統計量名)**, **par(パラメータ)**, **date(日付)**, **val(統計量)**} と定義し, 4 つの要素をまとめて **4 つ組**と呼ぶことにする. 各要素は, MuST コーパス中で, name, par, date, val タグが付与されている.

基本要素は, 文書中に明示的に示されていない場合がある. そのような場合の例として, 相対表現があげられる. 相対表現を抽出して他の情報と対応付けることができれば, 文書中に明示されていない情報を推論することができる.

例えば, 文書中に明示されている 4 つ組 (explicit な 4 つ組) として

{ ビール出荷量, アサヒ, 2007 年, 1 億 8824 万ケース }

が把握できているとする (図 1-1). 相対表現「前年比 0.1%増」を利用すれば,

$$\begin{aligned} \text{日付 } (date) : \text{前年} &= \text{2007 年 - 1 年} \\ &= \text{2006 年} \end{aligned}$$

$$\begin{aligned} \text{統計量 } (val) : 0.1\% &= \frac{\text{1 億 8824 万ケース}}{1 + \frac{0.1}{100}} \\ &= \text{1 億 8800 万ケース} \end{aligned}$$

という推論が可能である. その結果, 新らたに

{ ビール出荷量, アサヒ, 2006 年, 1 億 8800 万ケース }

というテキスト中に明示されていない 4 つ組 (implicit な 4 つ組) を獲得することができる.

3.2 相対表現の種類

今岡らは, 相対表現の機能に基づいて分類を行った (今岡裕貴他 2006). 彼らは, 相対表現には「日付を示す相対表現」と「統計量を示す相対表現」が存在すると報告している.

以下, それぞれの相対表現の機能について説明する.

3.2.1 日付を示す相対表現

新聞記事中に現れる日付は間接的に表現され, 相対表現によって示される場合が多い. 図 3-1 にその例を示す. この場合, 4 つ組 { ビール出荷量, アサヒ, 2007 年, 1 億 8824 万ケース } が得られる. さらに, 相対表現「前年」によって, 「日付 (date):前年」の存在が予測できる. 得られた 4 つ組を用いて

$$\begin{aligned} \text{日付}\ (date) : \text{前年} &= \text{2007 年 - 1 年} \\ &= \text{2006 年} \end{aligned}$$

が推論できる.

日付を示す相対表現は, 「今年」, 「昨日」, 「前年同月」, ··· などの形式で記述される.

2007年のアサヒのビール出荷量は **前年比0.1%増** の1億8824万ケースとなった。

図 3-1 日付と統計量の増減を表す相対表現の例

3.2.2 統計量を示す相対表現

新聞記事に現れる統計量は, 間接的に表現される場合が意外に多い. その多くは相対表現によって示されている. 本論文では, このような, 統計量を示す相対表現は, 表現が指し示す情報の種類によって, (1) 数値の増減を表す相対表現と, (2) 順位などの位置づけを表す相対表現, の 2 種類に大別できる.

(1) 数値の増減を示す相対表現

相対表現には, 時系列上, あるいは環境の変化に伴う統計量の増加・減少を示すものがある. 図 3-1 にその例を示す.

例文 1 では, 4 つ組として { ビール出荷量, アサヒ, 2007 年, 1 億 8824 万ケース } が得られる. また, 「前年比 0.1%増」という相対表現は, 日付と統計量の増加を示す. ここから,

$$\begin{aligned} \text{日付}\ (date) : \text{前年} &= \text{2007 年 - 1 年} \\ &= \text{2006 年} \end{aligned}$$

$$\begin{aligned} \text{統計量}\ (val) : 0.1\% &= \frac{\text{1 億 8824 万ケース}}{1 + \frac{0.1}{100}} \\ &= \text{1 億 8800 万ケース} \end{aligned}$$

が推論され, { ビール出荷量, アサヒ, 2006 年, 1 億 8800 万ケース } が得られる.

統計量の増減を示す相対表現は, 「... 年比～増」や「... に比べ～減」のような形式となり, 時間表現を示す相対表現と複合して記述される場合が多い.

(2) 順位などの位置づけを示す相対表現

(1) に対して, 同種の統計量の比較によって, その順位や相対的な位置づけを表す相対表現がある. 図 3-2 にその例を示す.

例文 1 では, { ビール出荷量, ϕ, 1999 年 6 月, 4200 万ケース } が得られる. ここで, 相対表現「過去最高」は, 関連付けられている統計量 (4200 万ケース) が, 比較される同種の統計量の中で最大であることを示す. したがって, 4 つ組で関連付けられている「日付 (date):1999 年 6 月」と比較し, これより過去の日付に関連付けられる統計量は「統計量 (val):4200 万ケース」以下であることが推測できる. 例文 2 では, 相対表現「1 位」, 「2 位」が利用でき, { ビールのメーカー別シェア, A 社, 12 月, 42.8%}, { ビールのメーカー別シェア, B 社, 12 月, 34.4%} が得られる. ここから, 以下の関係が推論できる.

A 社の統計量 (val) > B 社の統計量 (val)

位置づけを示す相対表現では, 「最高」や「最低」や「... 位」のような接辞を伴う場合が多い.

1. ビール出荷数量は 4200 万ケースとなり, 過去最高 となった.
2. 12 月のビールのメーカー別シェアでは 1 位A 社 (42.8%), 2 位B 社 (34.4%) だった.

図 3-2 順位や位置づけを表す相対表現の例

第4章

相対表現を利用した動向情報抽出

本章では, 相対表現を利用した動向情報抽出の基本的なモデルについて述べる. なお, 本章で述べる抽出モデルは, 7章での実験におけるベースラインとなる. 本抽出モデルは入力された統計量名 (クエリ) と関連がある explicit な 4 つ組を MuST コーパスと同様のタグが付与されたタグ付き文書から抽出し, 推論によって implicit な 4 つ組を生成する. 本モデルは, 以下のような 3 つの処理からなる.

(1) explicit な 4 つ組抽出

(2) 4 つ組の選択

(3) implicit な 4 つ組生成

各処理について説明する.

4.1 explicit な 4 つ組の抽出

explicit な 4 つ組抽出では, 抽出パターンを用いて, 4 つ組を構成する基本要素と相対表現を抽出する. さらに, 補完規則によって抽出パターンでは抽出できなかった要素を補完し, explicit な 4 つ組を抽出する.

4.1.1 基本要素抽出

抽出パターンを用いて, 4 つ組を構成する基本要素と相対表現を抽出する. 抽出パターンは, 今岡ら (今岡裕貴他 2006) の手法に基づき MuST コーパスにおける相対表現の出現パターンを人手で分析し, 抽象化することによって作成した. 図 4-1 に抽出パターンの例を示す.

本モデルはタグ付き文書を処理対象としているため, タグを手がかりとして, 4 つ組の各要素を特定することが可能となる.

Ex1.
<date> の <par> の <name> は <date> 比 <rel> 増の <val>

Ex2.
<name> は <date> 比 <rel> 減の <val>

図 4-1 抽出パターンの例

対象文書:
<date>2007 年 </date> の <par> アサヒ </par> の <name> ビール出荷量 </name> は <date> 前年 </date> 比 <rel>0.1%</rel> 増の <val>1 億 8824 万ケース </val> となった。

⇑

抽出パターン:
<date> の <par> の <name> は <date> 比 <rel> 増の <val>

⇓

name = ビール出荷量
par = アサヒ
date = 2007 年
val = 1 億 8824 万ケース

⇓

explicit な 4 つ組:
{ **ビール出荷量, アサヒ,2007 年,1 億 8824 万ケース**

図 4-2 基本要素の抽出

対象文書が入力されると, まず, 文書中に抽出パターンに適合する箇所が存在するかどうかを調べる. 規則に適合した場合, 適合箇所のタグに相当する文字列が対応する要素として抽出される. 図 4-2 の例では, 対象文書は図 4-1 の Ex1 に適合し, explicit な 4 つ組として (ビール出荷量, アサヒ,2007 年,1 億 8824 万ケース) が得られる.

4.1.2 不足要素の補完

抽出パターンだけでは, 4 つ組の構成要素抽出が不完全な場合があるため, 不足要素の補完処理を行う. 以下に, 各要素における補完対象となる要素を示す. なお, val 要素は補完に曖昧性が生じ性能を下げる原因となるため抽出パターンのみで抽出を行い, 補完を行わない.

name: 抽出パターンの適合箇所の前方, かつ, 最近傍の name 要素

par: 抽出パターンの適合箇所の前方, かつ, 同一文内の par 要素
同一文内で par 要素が無い場合, 4 つ組は par 要素を持たないとする.

date: 以下 3 つの条件を順に調べ, 適合する date 要素
(1) 抽出パターンの適合箇所の前方, かつ, 同一文内の date 要素
(2) 記事の冒頭から最も早く出現する date 要素
(3) 記事 ID の日付

例えば, 図 4-3 の例文において抽出パターン「<name> は <date> 比 <rel> 減の <val>」が適合すると name 要素として「パソコン出荷台数」, val 要素として「1414 万台」が獲得される. しかし, par 要素と date 要素は抽出パターンでは抽出されないため, 補完処理を行う. この場合 par 要素は抽出パターンの適合箇所と同一文中に無いのでパラメータ無し (ϕ) となる. date 要素は補完要素 (2) である 2007 年が補完される. その結果, 4 つ組 { パソコン出荷台数, ϕ, 2007 年, 1414 万台 } が得られる.

日本電子工業振興協会は <date>9 日 </date>、<date>**2007 年** </date> のパソコン国内実績を発表した。<name> パソコン出荷台数 </name> は <date> 前年 </date>比 <rel>1%</rel> 減の <val>1414 万台 </val> だった。

図 4-3 不足要素の補完例

4.2 4つ組選択

4 つ組選択では, 抽出した 4 つ組から入力の統計量名 (クエリ) と関連がある 4 つ組を選択する. クエリと 4 つ組の name 要素と par 要素を用いて同一性を判定することにより選択する. 選択手法は, 以下のとおりである.

step 1 獲得された 4 つ組の集合から 4 つ組を取り出し, name 要素および par 要素を形態素解析する. 名詞, 未知語および接頭詞を抽出し, それぞれに対して形態素の系列 N,P を生成する.

step 2 系列 P を系列 N の先頭に結合し系列 Q を生成する.

step 3 クエリと系列 Q の表層情報を用いて同一性判定を行う. 同一性の判定基準はクエリと系列 Q の関係に応じて二つの基準を設定した (図 4-4).

(a) クエリと系列 Q の形態素数が等しい場合
クエリと系列 Q が表層的に一致する場合, 同一であると判定する.

(b) 系列 Q がクエリから生成される suffix 集合または prefix 集合のどれかの要素と一致する場合
一致した suffix または prefix をクエリから除いて, 残った形態素列が抽出された 4 つ組の name 要素が存在する文中に出現する場合, 同一であると判定する.

step 4 獲得した全ての 4 つ組に対して同様の処理を繰り返す.

4.3 implicit な 4 つ組の生成

本モジュールは相対表現に対応した推論規則 (図 4-5) を用いて implicit な 4 つ組の生成する. 推論規則は, 予め人手によって作成した. 推論規則を explicit な 4 つ組の要素に適用し計算することによって implicit な 4 つ組が生成される.

例えば, 図 1 の Ex1 から以下の 4 つの基本要素が抽出され, explicit な 4 つ組が抽出されたとする.

$$
\begin{aligned}
name_{exp} &= \text{ビール出荷量} \\
par_{exp} &= \text{アサヒ} \\
date_{exp} &= \text{2007 年} \\
val_{exp} &= \text{1 億 8824 万ケース} \\
&\Downarrow \\
&\{\text{ビール出荷量, アサヒ, 2007 年,1 億 8824 万ケース}\}
\end{aligned}
$$

この場合相対表現「前年比 0.1%増」に対応した規則が選択され, $date_{exp}$ と val_{exp} に適用される. 以下のように計算を行い, implicit な 4 つ組が生成される.

$$
\begin{aligned}
date_{imp} &= \text{2007 年 - 1 年} \\
&= \text{2006 年}
\end{aligned}
$$

$$
\begin{aligned}
val_{imp} &= \frac{\text{1 億 8824 万ケース}}{1 + \frac{0.1}{100}} \\
&= \text{1 億 8800 万ケース} \\
&\Downarrow \\
&\{\text{ビール出荷量, アサヒ, 2006 年, 1 億 8800 万ケース}\}
\end{aligned}
$$

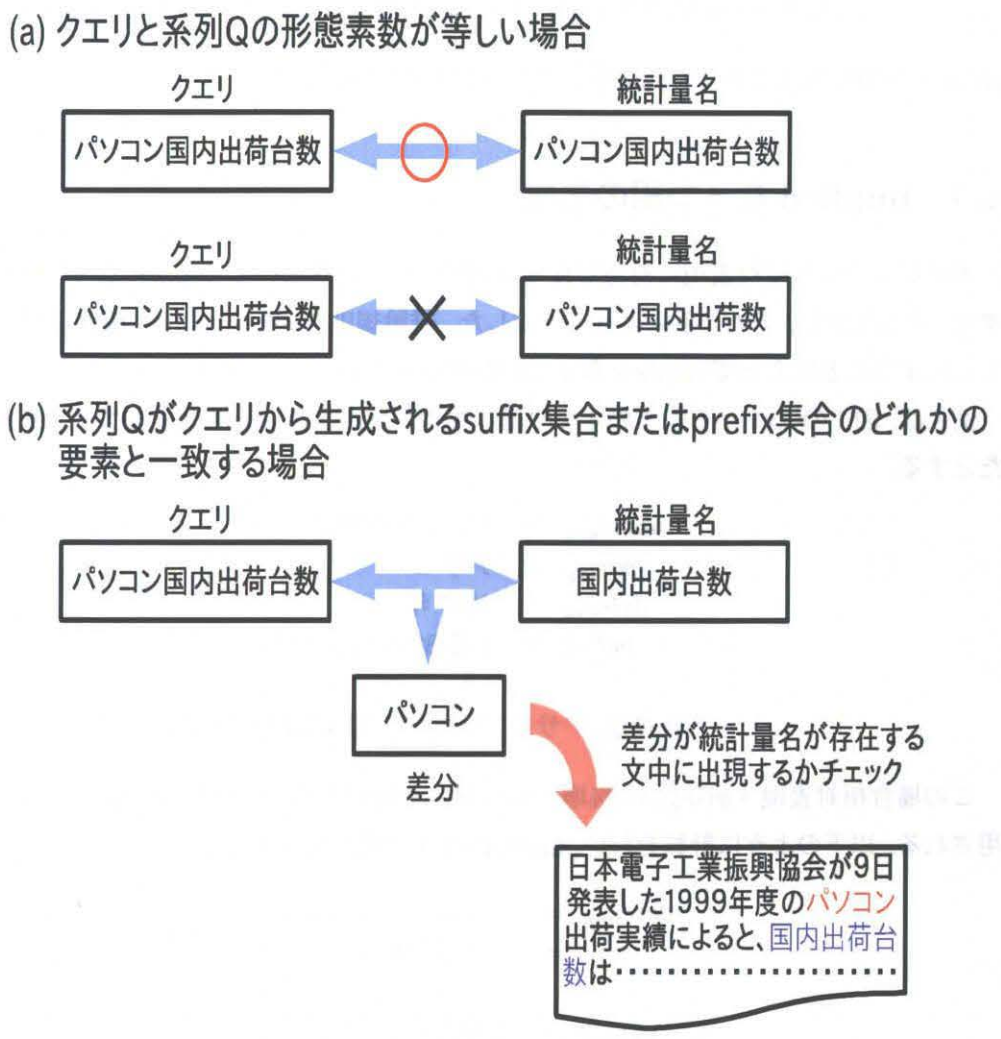


図 4-4 4 つ組選択の関連性判定の例

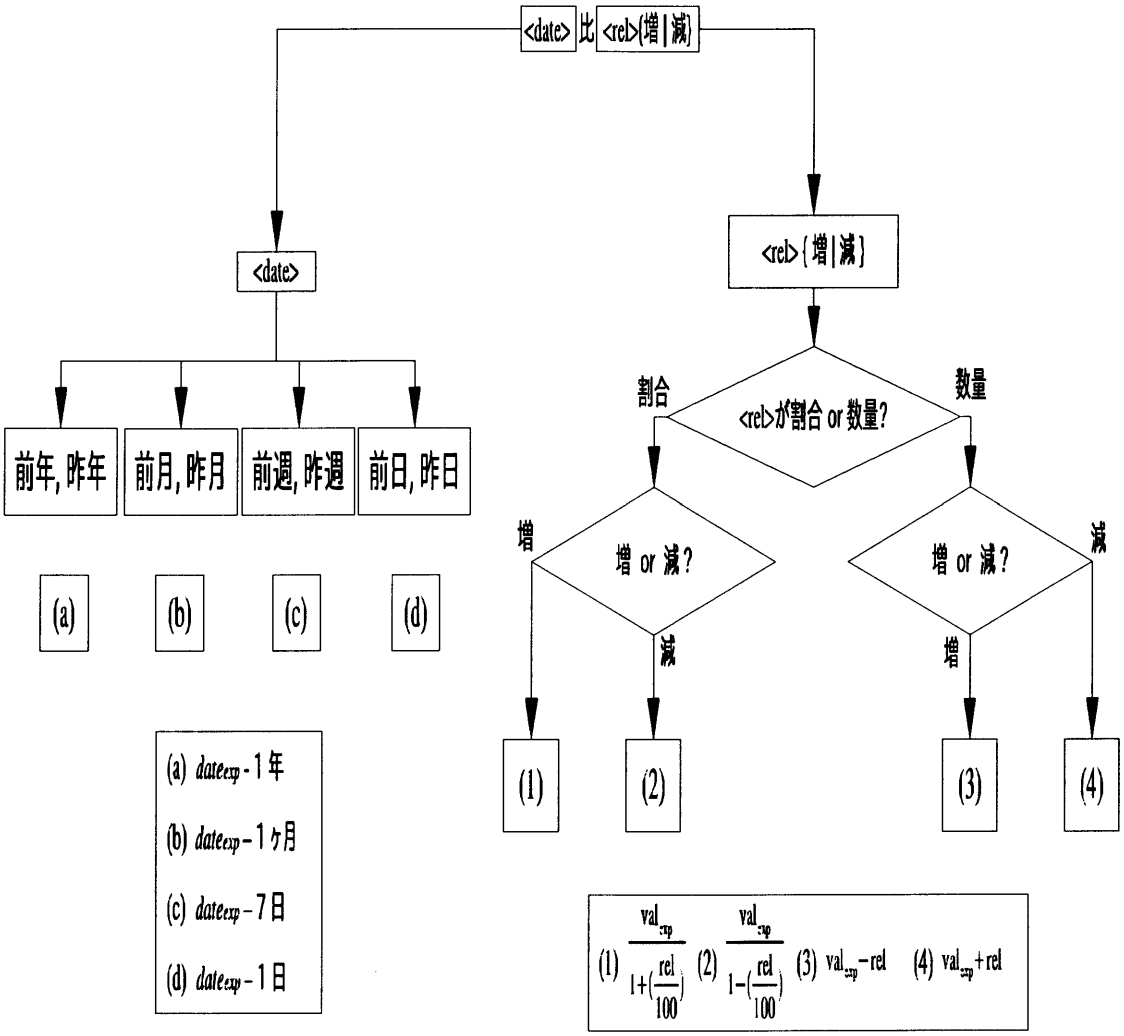

<date> 比 <rel>{増|減}
<date>
前年, 昨年
前月, 昨月
前週, 昨週
前日, 昨日
(a)
(b)
(c)
(d)
(a) $date_{exp}$ - 1年
(b) $date_{exp}$ - 1ヶ月
(c) $date_{exp}$ - 7日
(d) $date_{exp}$ - 1日
<rel>{増|減}
<rel>が割合 or 数量?
割合
数量
増 or 減?
増
減
増 or 減?
減
増
(1)
(2)
(3)
(4)
(1) $\frac{val_{exp}}{1+(\frac{rel}{100})}$ (2) $\frac{val_{exp}}{1-(\frac{rel}{100})}$ (3) $val_{exp}-rel$ (4) $val_{exp}+rel$

図 4-5 推論規則の例

4.4　4 つ組選択における課題

4.2 で述べた 4 つ組選択手法は, クエリと 4 つ組の name 要素と par 要素から生成される文字列との単純な表層的比較を行う. しかし, 文書中では, 同じ意味を表す name 要素でも様々な表記で現れる場合がある. 上記手法の単純な比較では, 同じ文字列のみを関連があると判定するため, このような, 表記が異なる文字列の関連性は判定できないため, 4 つ組の抽出漏れが生じると考えられる.

例えば, 「パソコン国内出荷台数」は「パソコン国内出荷数」, 「パーソナルコンピュータの国内出荷台数」, 「国内パソコン出荷台数」などのように複数の表記が存在する. クエリが「パソコン国内出荷台数」の場合は, 「パソコン国内出荷台数」のみを関連があると判定し, 「パソコン国内出荷数」, 「国内パソコン出荷台数」は関連がないと判定してしまう.

以降, これらの課題を解決する二つの選択手法をそれぞれ 5 章および 6 章で述べる.

第5章

4つ組選択手法1

本章では, 4.4 で述べた 4 つ組選択における課題を解決する第一の手法について述べる. 具体的には, 表記は異なるが同じ意味である「パソコン国内出荷台数」,「パソコン国内出荷数」,「パーソナルコンピュータの国内出荷台数」の関連性を判定する必要がある. 関連する 4 つ組を選択するためには, name 要素およびクエリの構成要素を検証する.

提案手法では, name 要素およびクエリは headword と specifier で構成されると仮定する. headword とは,「パソコン国内出荷台数」における「国内出荷台数」のような主辞を, specifier とは「パソコン」のような headword を修飾する部分を意味する.

構成要素の検証のためには, 以下のような 3 つの処理が適用される.

(1) suffix 集合の生成

(2) 構成要素の分割

(3) 構成要素の検証

構成要素の検証が成功すれば, name 要素はクエリと関連があるとする. そして, 関連がある name 要素を持つ 4 つ組を選択する.

表 5-1　name 要素またはクエリの構成要素の例

name element	specifier	headword
パソコン出荷台数	パソコン	出荷台数
ビール出荷数量	ビール	出荷数量
政党支持率	政党	支持率

構成要素を検証するために, EDR 電子化辞書[1] ,headword 辞書および関連語辞書を用いる. headword 辞書および関連語辞書は人手で構築された知識ベースである. headword 辞

[1]http://www.jsa.co.jp/EDR/index.html?

書はheadwordを検証するために使用し, 142語のheadwordが登録されている. 関連語辞書はspecifierを検証するために使用し, 54組の関係が登録されている. 以下, それぞれの処理について述べる.

5.1 suffix集合の生成:

クエリとname要素に対して, それぞれsuffix集合を生成する. まず, クエリを形態素解析し, 助詞を取り除き, 形態素の系列を作成する. 形態素の系列の各形態素から形態素の系列の末尾までの形態素の系列 (suffix) の集合を生成する. name要素に対しても, 同様の処理を行う. 例えば, 「パソコンの出荷台数」からは { 台数, 出荷台数, パソコン出荷台数 } のようなsuffix集合が生成される (図5-1).

Ex7.
パソコンの出荷台数
⇒ パソコン / 出荷 / 台数
⇒suffix: { 台数, 出荷台数, パソコン出荷台数 }

図 5-1 suffix集合の生成の例

5.2 構成要素の分割:

クエリおよびname要素をheadwordとspecifierに分割する.
まず, クエリに対して以下のようなステップが適用される.

step 1: 各suffixをheadword辞書のエントリと比較する.

step 2: エントリと一致した場合, 一致したsuffixがheadwordと認識される. 残りの部分がspecifierとして認識される.

step 3: 全てのsuffixが辞書のエントリと一致しなければ, 処理を終了する.

name要素に対しても同様のステップを適用する. name要素がheadwordだけで構成されている場合, par要素をspecifierとする.

例えば, 「パソコンの出荷台数」は「出荷台数 (headword)」と「パソコン (specifier)」に分割される (図5-2).

Ex8.
headword db:{ 出荷数量, 出荷台数, 支持率, ...}
suffix: { 台数, 出荷台数, パソコン出荷台数 }

suffix		headword 辞書
台数	⇒	"出荷数量" : miss
台数	⇒	"支持率" : miss
	⋮	
出荷台数	⇒	"出荷台数" : match
	⇓	
headword:		{ 出荷台数 (shipment volume)}
specifier:		{ パソコン (PC)}

図 5-2 構成要素の分割の例

5.3 構成要素の検証:

headword と specifier の関連性をそれぞれ検証する. The *name* element that passed both checks is determined to relate to the query word. 両検証を通った name 要素をクエリと関連があるとする.

(1) headword の検証:

step 1: クエリと name 要素の headword の ID を headword 辞書から取得し, ID を比較する.

step 2: ID が同じであれば, 次のステップへ進み, 同じでなければ処理を終了する.

(2) specifier の検証:

step 3: クエリと name 要素の specifier の ID を EDR 辞書から取得する. また, name 要素の上位概念の ID を EDR 辞書から取得する.

step 4: 以下のような場合, name 要素の specifier はクエリの specifire と関連があるとし, 処理を終了する.

(a) 両 specifier の ID が一致する場合 (Ex9).

(b) クエリの specifire の ID が name 要素の上位概念の ID と一致する場合 (Ex10).

step 5: 上記の判定に失敗した場合, 次のステップへ進む.

step 6: クエリと name 要素の specifier の ID を関連語辞書から取得する. また, name 要素の関連概念の ID を EDR 辞書から取得する.

step 7: 以下のような場合, name 要素の specifier はクエリの specifire と関連があるとし, 処理を終了する.

(a) 両 specifier の ID が一致する場合 (Ex11).

(b) クエリの specifire の ID が name 要素の関連概念の ID と一致する場合 (Ex12).

Ex9.
query: パソコン出荷台数
[パソコン ⇒3c677f]
name: パーソナルコンピュータの出荷台数
[パーソナルコンピュータ ⇒3c677f]
Ex10.
query: 政党支持率
[政党 ⇒0f95e0]
name: 支持率
par: 自民党
[自民党 ⇒1f7f26⇒0f95e0]
Ex11.
query: デジタルカメラ出荷台数
[デジタルカメラ ⇒1]
name: デジカメの出荷台数
[デジカメ ⇒1]
Ex12.
query: パソコンの出荷台数
[パソコン ⇒50]
name: 出荷台数
par: NEC [NEC⇒74⇒50]

図 5-3 specifier の検証の例

Ex13.
query: パソコン出荷台数
name: 出荷台数
sentence: **パソコン** 出荷金額は前年比 1.9%減の 1 兆 7360 億円で,
出荷台数 は同 4.2%増の **1249 万台** となった

図 5-4 文内共起に基づいた specifier の検証の例

name 要素が headword のみで構成されている場合, かつ, 4 つ組を構成する par 要素が無い場合, 文内共起に基づいて specifire を検証する. val 要素と相対表現が抽出された文中の抽出規則が適合した箇所より前方に対して, クエリの specifier と同じ文字列を探す. 発見できれば, name 要素をクエリと関連があるとする. 例えば, クエリの specifier「パソコン」は, val 要素「1249 万台」が抽出された文中に存在するので, name 要素はクエリと関連があるとする (図 5-4).

第6章

4つ組選択手法2

本章では, 4.4節で述べた4つ組選択における課題を解決する第二の手法について述べる. 本手法は, クエリの全ての組合せを生成することにより, 「パソコン国内出荷台数」と「国内パソコン出荷台数」のような, 構成する単語の順序が異なる場合でも, 関連があると判定することができる. 本手法では, まず, explicitな4つ組抽出の前処理として, クエリの解析とそれに基づき関連文書検索を行う. 本手法を4章の動向情報抽出に適用する場合, 処理の流れは以下のようになる.

(1) クエリ解析・文書検索

(2) explicitな4つ組の抽出

(3) 4つ組の選択

(4) implicitな4つ組の生成

以下, クエリ解析・文書検索と4つ組選択について述べる.

6.1 クエリ解析・文書検索

クエリ解析・文書検索では, 入力されたクエリを解析し, クエリと関連した文書を検索する. ここで, 「パソコン国内出荷台数」における「出荷台数」のような主辞をheadword, 「パソコン国内」のようなheadwordを修飾する部分をspecifierと呼ぶことにする. headwordには「出荷台数」や「出荷数」のように異表記がある. より多くのクエリと関連がある4つ組を抽出するためににheadwordの言い替え表現を用いる.

あらかじめMuSTコーパスから人手でheadwordを収集し, headword辞書を構築して用いた. 図6-1にクエリ解析・文書検索の概要を示す.

以下, 処理の流れについて説明する.

step 1 headword 辞書を参照し, 適合した headword をクエリの headword と置き換え, 新たなクエリを作成する.

step 2 各クエリそれぞれに対して, 形態素解析[1] を行い, 名詞・未知語・接頭詞を抽出し, クエリリストに登録する.

step 3 各クエリに対して全ての組合せの形態素の系列を作成する.

step 4 対象タグつき文書に対して作成した形態素の系列中の形態素で AND 検索を行う.

step 5 文書を獲得できた形態素の系列をクエリリストに登録する.

step1 から step3 の処理は, クエリのバリエーションを増やすため, 関連したより多くの 4 つ組を選択することができる.

例えば, クエリが「パソコン国内出荷台数」である場合, 新たに「国内パソコン出荷台数」や「パソコン国内出荷数」などバリエーションが生成され, それぞれに関して 4 つ組が選択できる.

6.2 4つ組の選択

4 つ組の選択には 6.1 節のクエリ解析で獲得したクエリリストの各エントリと 4 つ組を構成する name 要素,par 要素を用いる. 以下, 選択方法について詳述する.

step 1 獲得された 4 つ組の集合から 4 つ組を取り出し, name 要素および par 要素を形態素解析する. 名詞, 未知語および接頭詞を抽出し, それぞれに対して形態素の系列 N,P を生成する.

step 2 系列 P を系列 N の先頭に結合し系列 Q を生成する.

step 3 リストからエントリ (形態素数 A の系列 q) を取り出し, 系列 Q(形態素数 B) と要素の表層情報を用いて関連性を判定する. 判定は系列 q と系列 Q の形態素数によって判定基準が異なる.

(1) 形態素数 $A \leq$ 形態素数 B の場合
系列 A の全ての形態素が系列 B の形態素と一致する.

(2) 形態素数 $A >$ 形態素数 B の場合

[1]形態素解析器 ChaSen ver.2.3.3. http://chasen-legacy.sourceforge.jp/

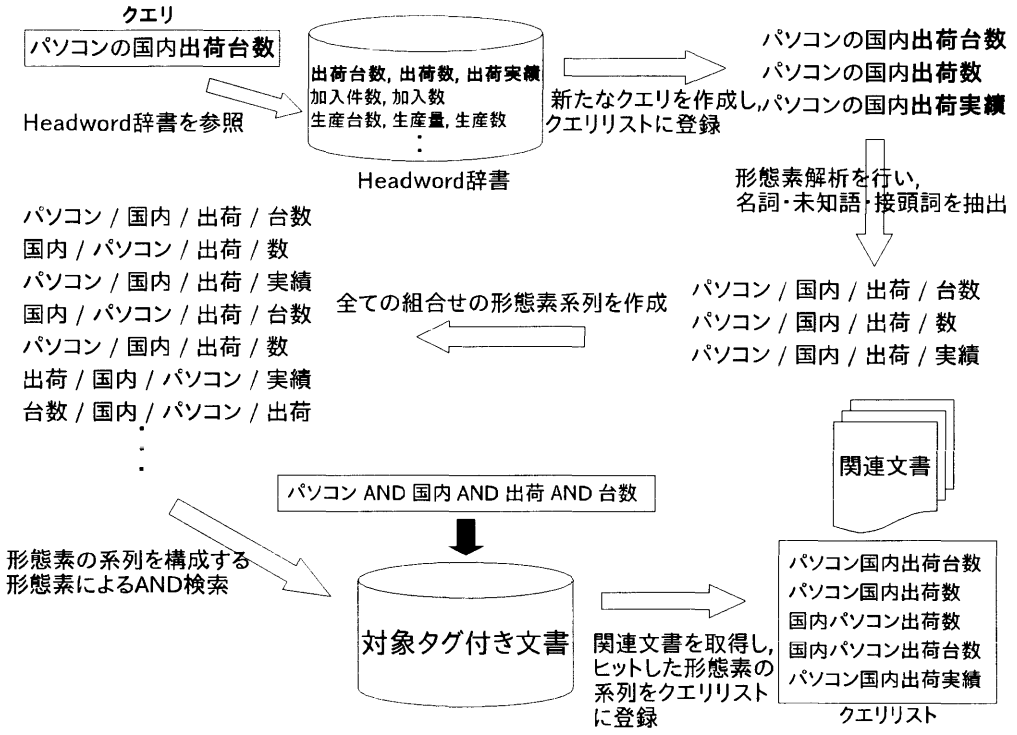


図 6-1　クエリ解析・文書検索の概要

- 系列 Q が系列 q のすべての形態素から末尾までの形態素の系列 S_i のいずれかに一致する, かつ, 系列 q と系列 q_i の差分の形態素の系列が name 要素の抽出文中または, その一つ前の文中に存在する.
- 系列 Q が系列 q のすべての形態素から先頭までの形態素の系列 T_i のいずれかに一致する, かつ, 系列 q と系列 q_i の差分の形態素の系列が name 要素の抽出文中または, その前方で最も近い name 要素が出現する文中に存在する.

step 4 関連がないと判定された場合, step3 に戻る.

step 5 全ての 4 つ組に対して同様の処理を繰り返す.

第7章

実験と評価

本章では, 前章までに述べた二つの提案手法の有効性を検証するために評価実験を行った. 本章では, 5章と6章で述べた4つ組選択手法を実装し, それぞれの手法を4章で述べた相対表現を利用した動向情報抽出に適用しシステムを構築した. それぞれのシステムの有効性を検証するためにベースラインシステムとの比較実験を行った. ベースラインシステムは, 4章で述べた相対表現を利用した動向情報抽出手法を実装したシステムを用いた.

7.1 実験

実験環境について述べる.

入力データとして, NTCIR MuST T2N subtask(Kato and Matsushita 2008) で用いられたクエリ (19個) を用いた. さらに, 2006年度の MuST コーパス中での相対表現の出現頻度が多いトピックから, クエリ (8個) を選び合計27個のクエリを用いた. また, 上記タスクで用いた1998年〜2001年の XML タグ付き毎日新聞100記事 (MuST コーパス) を使用した. さらに, 毎日新聞 (2000年から2003年) の43記事を使用した. 追加の43記事に対しては人手で MuST コーパスと同様のタグを付与した.

実験では, 以下の3つのシステムを用いた.

baseline: 4章で述べた相対表現を利用した動向情報抽出手法を実装したシステム. 本実験でのベースラインとなる.

system1: baseline の4つ組選択手法を5章で述べた手法に変更して実装したシステム.

system2: baseline の4つ組選択手法を6章で述べた手法に変更して実装したシステム. ただし, 関連性の判定条件を完全一致のみではなく, 表層的な包含関係を含めた場合.

system3: baseline の4つ組選択手法を6章で述べた手法に変更して実装したシステム. ただし, 関連性の判定条件を完全一致のみの場合.

実験の手順は, 以下のとおりである.

(1) クエリを一つ選択する.

(2) 選択したクエリに関する 4 つ組を含む記事を入力データのタグ付き新聞記事 143 記事から人手によって抽出する.

(3) クエリと (2) で抽出した記事をシステムに与え, クエリと関連する 4 つ組が正しく抽出されたかを評価する.

(4) すべてにクエリ (27 個) について同様の手順を繰り返す.

3 つのシステムは, 相対表現に関連した 4 つ組のみを抽出するので, 相対表現に関連した 4 つ組のみを評価対象とした. また, 本実験では explicit な 4 つ組のみを評価対象とした. 評価値として適合率, 再現率, F 値の 3 つの尺度を用いた. それぞれの評価値の定義を以下示す. なお, F 値の α は, 適合率と再現率のどちらを重要とするか度合を表すのパラメータである. 本実験では, 適合率, 再現率の両者を重要とするので, $\alpha = 0.5$ とした.

$$
\begin{aligned}
\text{適合率} &= \frac{\text{システムが抽出した正しい相対表現に関連する 4 つ組の数}}{\text{システムが抽出した相対表現に関連する 4 つ組の数}} \\
\text{再現率} &= \frac{\text{システムが抽出した正しい相対表現に関連する 4 つ組の数}}{\text{テストデータ中の相対表現に関連する 4 つ組の数}} \\
\text{F 値} &= \frac{1}{\alpha \frac{1}{\text{適合率}} + (1-\alpha)\frac{1}{\text{再現率}}}
\end{aligned}
$$

評価は MuST T2N subtask で配布された正解データを使用して評価を行った. 追加の 43 記事については, MuST T2N subtask の正解データに従って, 正解データを人手で作成し, 評価を行った.

表 7-1 有効性の評価

		baseline	system1	system2	system3
適合率	マイクロ平均	0.857(66/77)	0.677(84/124)	0.615(91/148)	0.848(84/99)
	マクロ平均	0.531	0.542	0.523	0.638
再現率	マイクロ平均	0.288(66/229)	0.367(84/229)	0.397(91/229)	0.367(84/229)
	マクロ平均	0.214	0.294	0.315	0.287
F 値	マイクロ平均	0.431	0.476	0.483	0.512
	マクロ平均	0.285	0.360	0.367	0.379

表 7-1 に評価結果を示す. baseline はマイクロ平均で適合率 0.857, 再現率 0.288, F 値 0.431 となった. マクロ平均では, 適合率 0.531, 再現率 0.214, F 値 0.285 となった. system1 はマイクロ平均で適合率 0.677, 再現率 0.367, F 値 0.476 となった. マクロ平均では, 適合

率 0.542, 再現率 0.294, F 値 0.360 となった. system2 はマイクロ平均で適合率 0.615, 再現率 0.397, F 値 0.483 となった. マクロ平均では, 適合率 0.523, 再現率 0.315, F 値 0.367 となった. system3 はマイクロ平均で適合率 0.848, 再現率 0.367, F 値 0.512 となった. マクロ平均では, 適合率 0.638, 再現率 0.287, F 値 0.379 となった.

7.2 考察

評価結果について考察する. system1 はマイクロ平均で再現率では 0.367 と baseline を上回った. このことは, headword と specifier を拡張することでより多くの 4 つ組を抽出できたことを示している. しかし, 適合率で 0.677 と baseline を下回った. この原因としては, 4 つ組選択手法において, 統計量名とクエリが一致した場合, パラメータを考慮していないことが考えられる. 例えば, クエリ「発泡酒の出荷数量」に対して, 「オリオンビールの発泡酒の出荷数量」に関する 4 つ組を抽出してしまった. 本実験での, 評価基準は MuST T2N subtask に基づいている. 上記の例の場合, MuST T2N subtask では, メーカ別でなく業界全体の発泡酒の出荷数量に関する 4 つ組のみを正解としているので, ノイズとなってしまい, 適合率が低下した.

system2 および system3 はマイクロ平均で再現率では 0.397, 0.367 と baseline を上回った. このことは, headword の言い替えを利用したことと組合せを生成したことが効果的に働いたと言える. 再現率に関しては, system2 の方が system3 よりも高い値となった. これは, system2 は関連性の判定基準において, クエリと完全一致ではなく, 表層的に包含関係にあるという基準に設定したためと考えられる. 実際に, system3 では適切に判定できなかったが system2 では判定できた例としては, クエリ「新設住宅着工戸数」に対して, 「新設住宅着工の総戸数」が挙げられる. この二つの文字列を比較すると, クエリの形態素系列の全ての形態素が「新設住宅着工の総戸数」という形態素系列の形態素と一致するため, 表層的な包含関係にあり, system2 では関連性有りと適切な判定ができた. しかし, system3 では, 完全一致では無いため関連性無しと誤った判定をしてしまった.

適合率に関しては, マイクロ平均で system2 は 0.615, system3 は 0.848 と system3 の方が高い値となった. これは, system2 では, 関連性判定の基準を表層的な包含関係にあるという基準に設定したことが挙げられる. 不正解例としては, クエリ「受験率」に対して「公民の受験率」を関連性ありと判定した例が挙げられる. 本実験での評価基準である MuST T2N subtask では, クエリ「受験率」に対してはセンター試験の全体の受験率のみを正解としており, 科目別の受験率などは正解には含まれていなかったため, 上記の例は不正解となり, 適合率の低下の原因となったと考えられる. しかし, 上記の例は必ずしも不正解とは言い切れない. なぜなら, ユーザの意図やニーズによっては科目別の受験率なども必要な情

報となりうることがあると考えられるからである.

3システムのうちF値ではsystem2が0.512と最も高かった. system2のマイクロ平均とマクロ平均を比較すると, 適合率, 再現率, F値の全てにおいて開きが生じている. これは, システムの性能に偏りがあることが示唆される. 再現率は0.367となった. これは, システムは4つ組を十分抽出しきれていないことがわかる. 表7-2にsystem2における正解を抽出できなかった原因の内訳を示す. 以下, 各原因について考察する.

表7-2 正解を抽出できなかった原因の内訳

検索ミス	パターンマッチ	不足要素補完	同一性判定	合計
25(18%)	41(28%)	29(20%)	50(34%)	145

検索ミスについて考察する. システムは基本的にはクエリを構成する形態素(名詞・未知語・接頭詞)の単純なAND検索を行っている. したがって, 正解を含むにもかかわらず, クエリを構成する全ての形態素が含まれないため, 検索することができない文書があった. 例えば, クエリが「携帯電話とPHSの合計加入台数」のような場合, 文書中で4つ組を構成するname要素が「携帯電話とPHSの加入台数」のように表現されており, 「合計」という形態素が文書中に存在しなかったことから, 正解を含む文書を検索できないことがあった. これに対応するには, 検索語を削除するなどして検索条件を緩めて再検索することが考えられる. そのような場合, 削除する形態素や検索を繰り返す条件などを考慮する必要がある. なぜなら, 関係のない文書を検索する恐れが生じ, 処理効率が悪くなると考えられるからである. 効率よく正解を含む文書を検索するためには, クエリを構成するそれぞれの形態素に対して$tf \cdot idf$のような頻度に基づいた重みをつけ, 重みに応じて削除する形態素の優先度を決定することが考えられる. 検索に失敗した他の例として, クエリが「デジタルカメラ国内出荷台数」のような場合, 文書中では「デジカメ国内出荷台数」のように「デジタルカメラ」が「デジカメ」のように言い替えられて表現されていたため, 検索に失敗した. これに対応するには, 5で述べたspecifier拡張手法のように, specifierの言い替え表現を利用することで対応可能であると考えられる.

パターンマッチについて考察する. 要因の一つは, パターンが不足していたことが挙げられる. 「センター試験」トピックでは, 「<name>は<date>より<rel>多い<val>」のようなパターンが不足していた. 「センター試験」トピックはパターンの作成に用いたMuSTコーパス中には無いトピックであったことから, MuSTコーパスの網羅性が十分でないことが示唆される. また, 現在は人手での作成となっているため, 多くのパターンを作るにはコストが掛かると考えられる. パターンの自動作成, または自動で拡張するなどの対

策が今後の課題として挙げられる.

不足要素の補完について考察する. 不足要素の補完の失敗で顕著に見られたのは, date要素の補完であった. クエリ「内閣支持率」や「自民党支持率」の「政治動向」トピックでは, 基本要素の抽出においてパターンが不足していたことから, 補完の失敗に継った. 例えば, 「<date> 昨年 1 2月 </date> の <date> 前回 </date> 面接調査に比べ <rel> 9ポイント </rel> 増」のような相対表現を誤って「<date> 前回 </date> 面接調査に比べ <rel> 9ポイント </rel> 増」と認識してしまったことから, 「<date> 昨年 1 2月 </date>」が補完対象の候補となって, 誤って補完された.

4つ組選択における同一性判定について考察する. 同一性の判定誤りの原因は, 記事中でしばしば出現する「出荷台数」のようなspecifierが省略された表現や「携帯電話」のようなheadwordが省略された表現に対応しきれなかったことが挙げられる. このような場合, システムは, クエリとname要素の差分の形態素が4つ組のname要素が存在する文, または, そのname要素の一つ前のname要素が出現する文中に出現すれば同一と判定していた. 対象範囲を広げることで, 正解数は増えると予想されるが, 同時に判定の誤る回数も増えるため, 最善の策とは言いがたい. 適切な判定を行うには, 省略されている形態素を補完して, 完全な統計量名を生成することが考えられる. これには, 森らの機械学習を用いた統計量名抽出手法 (森辰則, 藤岡篤史, 村田一郎 2008)(森辰則 上野史紀 2008) が応用できると考えられる.

第8章

結論

本論文では, 新聞記事からの動向情報抽出を目的として, 2 種類の 4 つ組選択手法を提案し, それぞれを組み込んだ相対表現を利用した動向情報抽出手法を提案した. 第一の 4 つ組選択手法は, 4 つ組の選択において電子化辞書と人手で作成した同義語辞書および関連語辞書を用いて選択するという特徴がある. 第二の 4 つ組選択手法は, 4 つ組の選択において入力となる統計量名 (クエリ) に対して, 全ての形態素の組合せを作成し, さらに同義語辞書を用いてクエリのバリエーションを増やして選択するという特徴がある.

新聞記事を用いて評価実験したところ, 第一の 4 つ組選択手法を組み込んだ手法はマイクロ平均で適合率 0.677, 再現率 0.367, F 値 0.476 という性能が得られた. 第二の 4 つ組選択手法を組み込んだ手法は関連性判定を表層的な包含関係とした場合, マイクロ平均で適合率 0.615, 再現率 0.397, F 値 0.483 という性能が得られた. 関連性判定を完全一致とした場合, マイクロ平均で適合率 0.848, 再現率 0.367, F 値 0.512 という性能が得られた. このことから, 提案する相対表現に基づいた動向情報抽出の有効性が明らかとなった.

今後の課題としては, タグ無しの文書に対応することが挙げられる. 本手法は, 文書タグ付き文書を前提としているため, 処理が可能となる文書が限られてしまうからである. タグ無しの文書に対応するための一つの方法として, 固有表現抽出器 (渡辺一郎, 桝井文人, 福本淳一 2004) に本手法を実装することが考えられる.

謝辞

本論文に関する研究を進め, 論文を完成させるにあたり, 本当に多くご指導とご支援を賜りました, 桝井文人助教, 井須尚紀教授, 河合敦夫准教授に深く感謝致します. 本研究についての助言, ならびにMuSTコーパスを提供していただいた, MuSTワークショップオーガナイザーである東京大学大学院総合文化研究化言語情報科学専攻の加藤恒昭准教授に深く感謝致します. 書き換え変換モジュールについて助言をしていただきました三重大学総合情報処理センター三橋一郎助教に深く感謝致します. 貴重な時間をさいて本論文を査読していただきました若林哲史准教授に深く感謝致します. いろいろと便宜を図っていただきました田中みゆき事務官, 吉永みゆき事務官に深く感謝致します. 最後に, 楽しい研究生活を共にした, 研究室の皆様に深く感謝致します.

本研究は科研費（20500833）の助成を受けたものである.

文献

今岡裕貴, 桝井文人, 河合敦夫, 井須尚紀 (2006). "動向情報抽出における相対表現の利用効果に関する考察." 日本知能情報ファジィ学会誌, **18** (5), 735–744.

加藤恒昭 松下光範 (2006). "情報編纂 (Information Compilation) の基盤技術." 人工知能学会第 20 回全国大会講演論文集, 1D3–2.

Kato, T. and Matsushita, M. (2008). "Overview of MuST at the NTCIR-7 Workshop – Challenges to Multi-modal Summarization for Trend Information –." In *Proceedings of NTCIR-7 Workshop Meeting*, pp. 475–488.

KATO, T., MATSUSHITA, M., and KANDO, N. (2005). "A Workshop on Multimodal Summarization for Trend Information." In *Proceedings of NTCIR-5 Workshop Meeting*, pp. 556–563.

森辰則, 藤岡篤史, 村田一郎 (2008). "動向情報編纂のためのテキストからの統計量表現の自動抽出." 人工知能学会論文誌, **23** (5), 310–318.

難波英嗣, 国政美伸, 福島志穂, 相沢輝昭, 奥村学 (2005). "文書横断文間関係を考慮した動向情報の抽出と可視化." 情報処理学会研究報告, 05-NL-168, pp. 67–74.

NANBA, H., OKUDA, N., and OKUMURA, M. (2007). "Extraction and Visualization of Trend Information from Newspaper Articles and Blogs." In *Proceedings of NTCIR-6 Workshop Meeting*, pp. 243–248.

Takama, Y. (2007). "Visualization of Earthquake Trend Information from MuST Corpus." In *Proceedings of NTCIR-6 Workshop Meeting*, pp. 249–255.

Yoshida, M., Sugiura, T., Hirokawa, T., Yamada, K., Masuda, H., and Nakagawa, H. (2008). "TDU Systems for MuST: Attribute Name Extraction, Text-Based Stock Price Analysis, and Automatic Graph Generation." In *Proceedings of NTCIR-7 Workshop Meeting*, pp. 520–527.

松下光範 加藤恒昭 (2007). “Elucignage: 探索的データ分析のための動向情報可視化インタフェース.” 動向情報の要約と可視化に関するワークショップ第二回成果進捗報告会予稿集, pp. 17–18.

森辰則 上野史紀 (2008). “動向情報編纂のためのテキストからの統計量表現の自動抽出.” 人工知能学会第 20 回全国大会講演論文集, 3K3–06.

渡辺一郎, 桝井文人, 福本淳一 (2004). “固有表現抽出ツール NExT の精緻化とユーザビリティの向上.” 第 10 回言語処理学会年次大会発表論文集, pp. 413–415.

AppendixA

実験結果の詳細

表 A-1 にテストデータにおける各クエリ毎の正解の 4 つ組の数を示す. baseline(表 A-2), system1(表 A-3), system2(表 A-4), system3(表 A-5) の実験結果の詳細を示す.

表 A-1 テストデータ中の正解の4つ組数

	4つ組数
レギュラーガソリンの全国平均店頭価格	6
ドバイ原油価格	5
周辺機器を含むパソコン国内出荷金額	4
パソコン国内出荷台数	18
携帯電話の加入者数	9
ＰＨＳの加入台数	6
携帯電話とＰＨＳの合計加入台数	4
内閣支持率	10
内閣不支持率	11
自民党政党支持率	9
民主党政党支持率	8
鉱工業生産指数	24
鉱工業出荷指数	9
鉱工業在庫指数	10
デジカメの国内出荷台数	8
デジカメの国内出荷額	4
志願者数	12
志願倍率	6
受験率	4
ビール・発泡酒の出荷数量合計	6
ビール・発泡酒総市場における発泡酒の割合	4
ビールの出荷数量	8
発泡酒の出荷数量	8
サラリーマン世帯の消費支出	7
全世帯の消費支出	4
新設住宅着工戸数	13
日経平均株価	12
合計	229

表 A-2　baseline の結果

	抽出数	正解数	適合率	再現率	F 値
レギュラーガソリンの全国平均店頭価格	0	0	0.000	0.000	0.000
ドバイ原油価格	1	1	1.000	0.200	0.333
周辺機器を含むパソコン国内出荷金額	0	0	0.000	0.000	0.000
パソコン国内出荷台数	3	3	1.000	0.167	0.286
携帯電話の加入者数	0	0	0.000	0.000	0.000
ＰＨＳの加入台数	0	0	0.000	0.000	0.000
携帯電話とＰＨＳの合計加入台数	0	0	0.000	0.000	0.000
内閣支持率	3	2	0.667	0.200	0.308
内閣不支持率	0	0	0.000	0.000	0.000
自民党政党支持率	7	3	0.429	0.333	0.375
民主党政党支持率	7	3	0.429	0.375	0.400
鉱工業生産指数	20	20	1.000	0.833	0.909
鉱工業出荷指数	5	5	1.000	0.556	0.714
鉱工業在庫指数	2	2	1.000	0.200	0.333
デジカメの国内出荷台数	0	0	0.000	0.000	0.000
デジカメの国内出荷額	0	0	0.000	0.000	0.000
志願者数	1	1	1.000	0.083	0.154
志願倍率	2	2	1.000	0.333	0.500
受験率	0	0	0.000	0.000	0.000
ビール・発泡酒の出荷数量合計	0	0	0.000	0.000	0.000
ビール・発泡酒総市場における発泡酒の割合	0	0	0.000	0.000	0.000
ビールの出荷数量	2	2	1.000	0.250	0.400
発泡酒の出荷数量	3	3	1.000	0.375	0.545
サラリーマン世帯の消費支出	3	3	1.000	0.429	0.600
全世帯の消費支出	1	1	1.000	0.250	0.400
新設住宅着工戸数	7	7	1.000	0.538	0.700
日経平均株価	10	8	0.800	0.667	0.727
マイクロ平均	77	66	0.857	0.288	0.431
マクロ平均			0.531	0.214	0.285

表 A-3 system1 の結果

	抽出数	正解数	適合率	再現率	F 値
レギュラーガソリンの全国平均店頭価格	0	0	0.000	0.000	0.000
ドバイ原油価格	0	0	0.000	0.000	0.000
周辺機器を含むパソコン国内出荷金額	0	0	0.000	0.000	0.000
パソコン国内出荷台数	7	6	0.857	0.333	0.480
携帯電話の加入者数	4	4	1.000	0.444	0.615
ＰＨＳの加入台数	2	2	1.000	0.333	0.500
携帯電話とＰＨＳの合計加入台数	0	0	0.000	0.000	0.000
内閣支持率	10	3	0.300	0.300	0.300
内閣不支持率	0	0	0.000	0.000	0.000
自民党政党支持率	8	3	0.375	0.333	0.353
民主党政党支持率	8	3	0.375	0.375	0.375
鉱工業生産指数	20	20	1.000	0.833	0.909
鉱工業出荷指数	5	5	1.000	0.556	0.714
鉱工業在庫指数	2	2	1.000	0.200	0.333
デジカメの国内出荷台数	3	2	0.667	0.250	0.364
デジカメの国内出荷額	3	1	0.333	0.250	0.286
志願者数	6	5	0.833	0.417	0.556
志願倍率	5	4	0.800	0.667	0.727
受験率	2	1	0.500	0.250	0.333
ビール・発泡酒の出荷数量合計	0	0	0.000	0.000	0.000
ビール・発泡酒総市場における発泡酒の割合	0	0	0.000	0.000	0.000
ビールの出荷数量	2	2	1.000	0.250	0.400
発泡酒の出荷数量	3	2	0.667	0.250	0.364
サラリーマン世帯の消費支出	4	3	0.750	0.429	0.545
全世帯の消費支出	1	1	1.000	0.250	0.400
新設住宅着工戸数	19	7	0.368	0.538	0.438
日経平均株価	10	8	0.800	0.667	0.727
マイクロ平均	124	84	0.677	0.367	0.476
マクロ平均			0.542	0.294	0.360

表 A-4　system2 の結果

	抽出数	正解数	適合率	再現率	F 値
レギュラーガソリンの全国平均店頭価格	0	0	0.000	0.000	0.000
ドバイ原油価格	1	1	1.000	0.200	0.333
周辺機器を含むパソコン国内出荷金額	0	0	0.000	0.000	0.000
パソコン国内出荷台数	10	10	1.000	0.556	0.714
携帯電話の加入者数	7	6	0.857	0.667	0.750
ＰＨＳの加入台数	4	3	0.750	0.500	0.600
携帯電話とＰＨＳの合計加入台数	0	0	0.000	0.000	0.000
内閣支持率	10	3	0.300	0.300	0.300
内閣不支持率	6	2	0.333	0.182	0.235
自民党政党支持率	8	3	0.375	0.333	0.353
民主党政党支持率	8	3	0.375	0.375	0.375
鉱工業生産指数	20	20	1.000	0.833	0.909
鉱工業出荷指数	5	5	1.000	0.556	0.714
鉱工業在庫指数	2	2	1.000	0.200	0.333
デジカメの国内出荷台数	0	0	0.000	0.000	0.000
デジカメの国内出荷額	1	1	1.000	0.250	0.400
志願者数	4	3	0.750	0.250	0.375
志願倍率	6	2	0.333	0.333	0.333
受験率	3	1	0.333	0.250	0.286
ビール・発泡酒の出荷数量合計	0	0	0.000	0.000	0.000
ビール・発泡酒総市場における発泡酒の割合	0	0	0.000	0.000	0.000
ビールの出荷数量	6	3	0.500	0.375	0.429
発泡酒の出荷数量	6	3	0.500	0.375	0.429
サラリーマン世帯の消費支出	5	3	0.600	0.429	0.500
全世帯の消費支出	1	1	1.000	0.250	0.400
新設住宅着工戸数	25	8	0.320	0.615	0.421
日経平均株価	10	8	0.800	0.667	0.727
マイクロ平均	148	91	0.615	0.397	0.483
マクロ平均			0.523	0.315	0.367

表 A-5 system3 の結果

	抽出数	正解数	適合率	再現率	F 値
レギュラーガソリンの全国平均店頭価格	0	0	0.000	0.000	0.000
ドバイ原油価格	1	1	1.000	0.200	0.333
周辺機器を含むパソコン国内出荷金額	0	0	0.000	0.000	0.000
パソコン国内出荷台数	8	8	1.000	0.444	0.615
携帯電話の加入者数	5	5	0.000	0.556	0.000
ＰＨＳの加入台数	3	3	0.000	0.500	0.000
携帯電話とＰＨＳの合計加入台数	0	0	0.000	0.000	0.000
内閣支持率	3	2	0.667	0.200	0.308
内閣不支持率	2	0	0.000	0.000	0.000
自民党政党支持率	8	3	0.375	0.333	0.353
民主党政党支持率	8	3	0.375	0.375	0.375
鉱工業生産指数	20	20	1.000	0.833	0.909
鉱工業出荷指数	5	5	1.000	0.556	0.714
鉱工業在庫指数	2	2	1.000	0.200	0.333
デジカメの国内出荷台数	0	0	0.000	0.000	0.000
デジカメの国内出荷額	1	1	0.000	0.250	0.000
志願者数	4	4	1.000	0.333	0.500
志願倍率	2	2	1.000	0.333	0.500
受験率	0	0	0.000	0.000	0.000
ビール・発泡酒の出荷数量合計	0	0	0.000	0.000	0.000
ビール・発泡酒総市場における発泡酒の割合	0	0	0.000	0.000	0.000
ビールの出荷数量	3	3	1.000	0.375	0.545
発泡酒の出荷数量	3	3	1.000	0.375	0.545
サラリーマン世帯の消費支出	3	3	1.000	0.429	0.600
全世帯の消費支出	1	1	1.000	0.250	0.400
新設住宅着工戸数	7	7	1.000	0.538	0.700
日経平均株価	10	8	0.800	0.667	0.727
マイクロ平均	99	84	0.848	0.367	0.512
マクロ平均			0.527	0.287	0.313

AppendixB

書き換え知識変換モジュール

書き換え知識変換モジュールは図 B-3 に示す構文図式に従って書かれた書き換え知識を正規表現を用いた perl プログラムに変換するモジュールである. 図 B-1 に書き換え知識変換モジュールの入力と出力の例を示す. 書き換え知識は書き換え規則と展開規則の集合から構成される.

書き換え規則とは, 図 B-1 の (a) のような形式をする. 「==>」の左右にパタンが記述される. これ以下, 「==>」の左の文字列を左辺, 右の文字列を右辺と呼ぶこととする. 書き換え規則は, 左辺とマッチする文字列があれば, 右辺のように書き換えるということを指定する.

書き換え規則の左辺は以下の要素が並んだ文字列である.

<\tag>	<tag>...</tag> のような tag というタグで囲まれた部分 (以降, タグ要素と呼ぶ) の略記.
<\tag attr=val>	tag というタグ要素で, かつ, attr 属性の値が val であるものを示す. val の位置に*があった場合 (attr=*の場合), 任意の attr 属性の値が存在していることを指定する. val は, XML の仕様に沿って, ""で囲まれた文字列とする.
<\tag attr!=val>	tag というタグ要素で, かつ, attr 属性の値が val でないものを示す. val の位置に*があった場合 (attr!=*の場合), そのタグ要素に attr 属性が存在していないことを示す.
' ' で囲まれた文字列	終端記号であり, そのままの文字列を示す.
それ以外の全角文字列	非終端記号であり, 展開規則の指定に沿って展開される.
(...)	グルーピング. ... の部分をグループとして指定する.
(.+?)	タグを含まない文字列で, 1 文字以上の最短文字列とマッチする. (このままでは, グループでない, グループにする場合はもう一度括弧で囲む)
(.*?)	タグを含まない文字列で, 0 文字以上の最短文字列とマッチする. (このままでは, グループでない, グループにする場合はもう一度括弧で囲む)

$	句読点あるいは unit の要素の終端とマッチする.
(?:...)	後で参照されないグルーピング.
{...|...}	選択. {"abc"—"def"} 等で abc もしくは def とマッチする.
文字列?	指定された文字列 (' ' で囲まれた終端記号または" "なしの非終端記号) の 0 回または 1 回の繰り返しを指定する. ?のスコープが文字列全体であることに注意.

書き換え規則の右辺は以下のように$n を幾つかのタグが挟んだ形式をしている (左右の開始 tag と終了 tag は入れ子の関係で対応していなければならない).

<tag1...>...<tagN...>$n</tagN>...<tag1>

ここで, $n(n=0,1,2...) は左辺において, <tag> 等で指定されたタグ要素かグルーピングによって指定された部分とマッチした文字列で, パタンの左側から$1, $2, $3, ... となる. $0 は左辺でマッチした全体を示す. 右辺の各行は, それら ($n) の部分を指定したタグ要素とすることを意味する. 右辺の <tag...> において, ... の部分には以下のものが記述でき, 属性の値の引き継ぎを行う.

$n	左辺において$n にタグ要素がマッチしている必要があり, そのマッチしたタグ要素の全ての属性とその値がここに引き継がれることを指定する.
attr=$n.attr1	左辺において$n にタグ要素がマッチしている必要があり, そのマッチしたタグ要素の attr1 属性の値が attr 属性の値となることを指定している. ここで, attr1 が text の場合は, マッチしたタグ要素のタグで囲まれている文字列 (タグは除く) が attr 属性の値となることを指定している.
attr=val	val は" "で囲まれた文字列. 左辺からの値の引き継ぎでなく, 値を指定する.

展開規則とは, 図 B-1 の (b) のような形式をする. 「::=」の左の部分の文字列は書き換え規則の左辺で用いられる文字列である. 展開規則はそれらの文字列をより具体的な同クラスの要素の集合に変換することを指定する.

例えば, 文書中に図 B-2 の書き換え前のような文字列が存在した場合, 図 B-1 の書き換え知識によって, 図 B-2 の書き換え後のように書き換えられる.

入力: 書き換え知識

(a) 書き換え規則
<\date>"に"<\val>"まで" 上昇句 "した" ==>
<ipc date=$1.abs type0="dur_eq", val=$2.text type1="upward"> $0
</ipc>

(b) 展開規則
上昇句 ::= {"上昇" | "上げ" | "増加"}

出力: perl プログラム

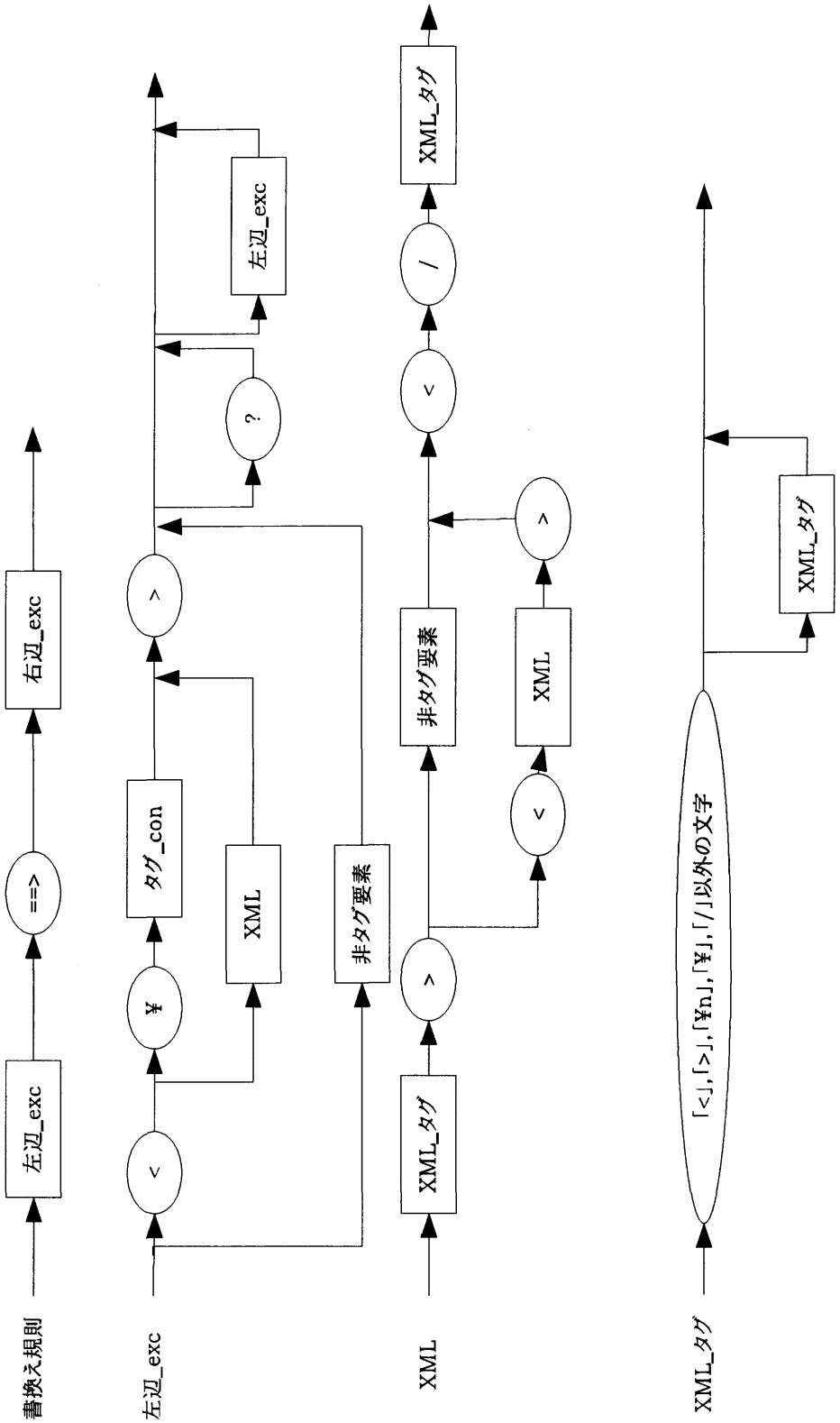
```
if ($text =~ /^(<date[^<>]*?abs=("[^<>]*?")[^<>]*?>[^<>]+?
<\/date>)(に)(<val[^<>]*?>([^<>]+?)<\/val>)(まで)((?:上昇|上げ|
増加))(した)/g){
    $text =~ s/(<date[^<>]*?abs=("[^<>]*?")[^<>]*?>[^<>]+?
    <\/date>)(に)(<val[^<>]*?>([^<>]+?)<\/val>)(まで)((?:上昇|
上げ|増加))(した)/$1$3$4$6$7$8/;
    $text =~ s/(<date[^<>]*?abs=("[^<>]*?")[^<>]*?>[^<>]+?
    <\/date>)(に)(<val[^<>]*?>([^<>]+?)<\/val>)(まで)((?:上昇|
上げ|増加))(した)/val="$5" type1="upward"> $4 <\/ipc>/;
}
```

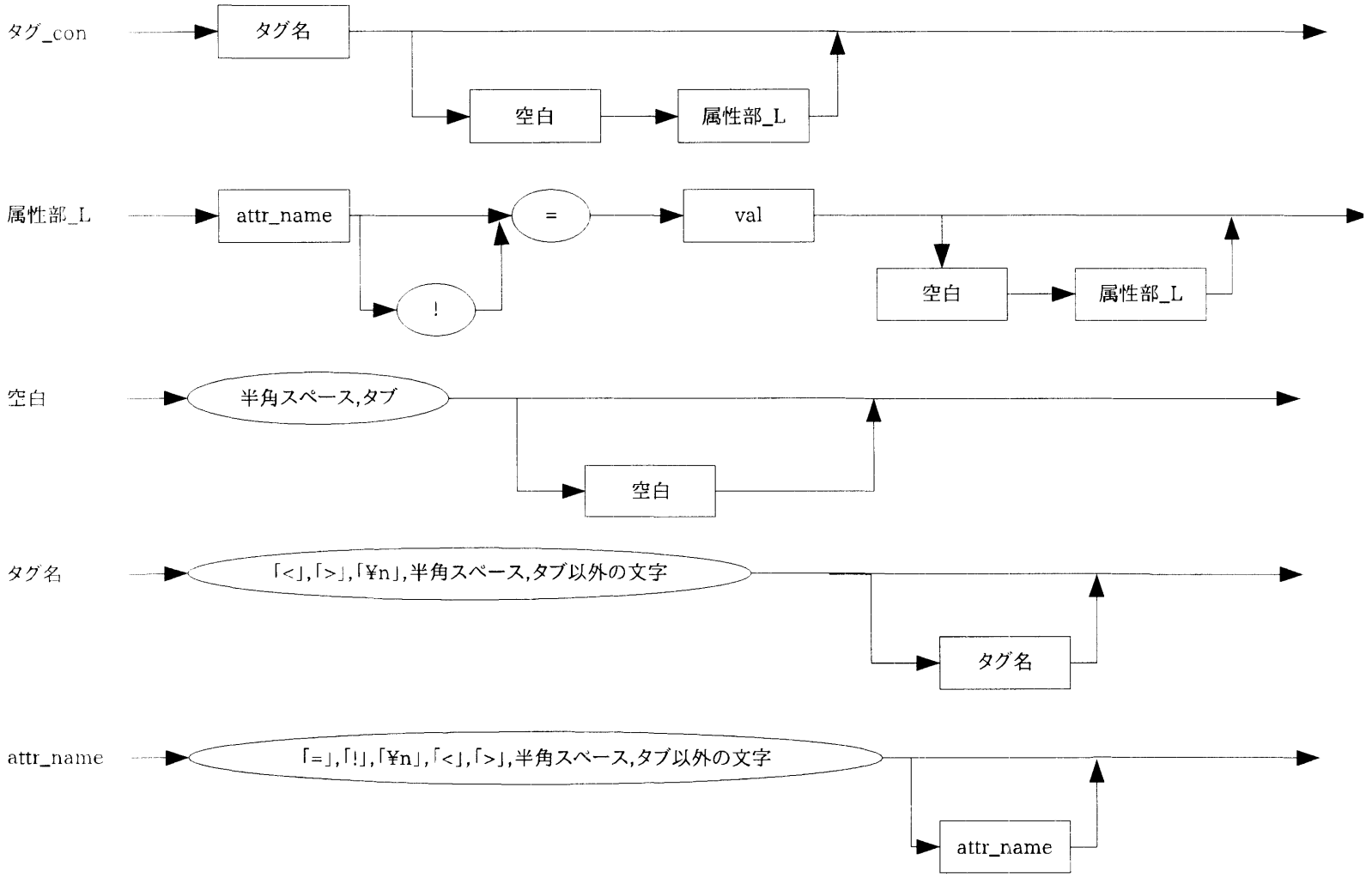
図 B-1　書き換え知識変換モジュールの入力と出力の例

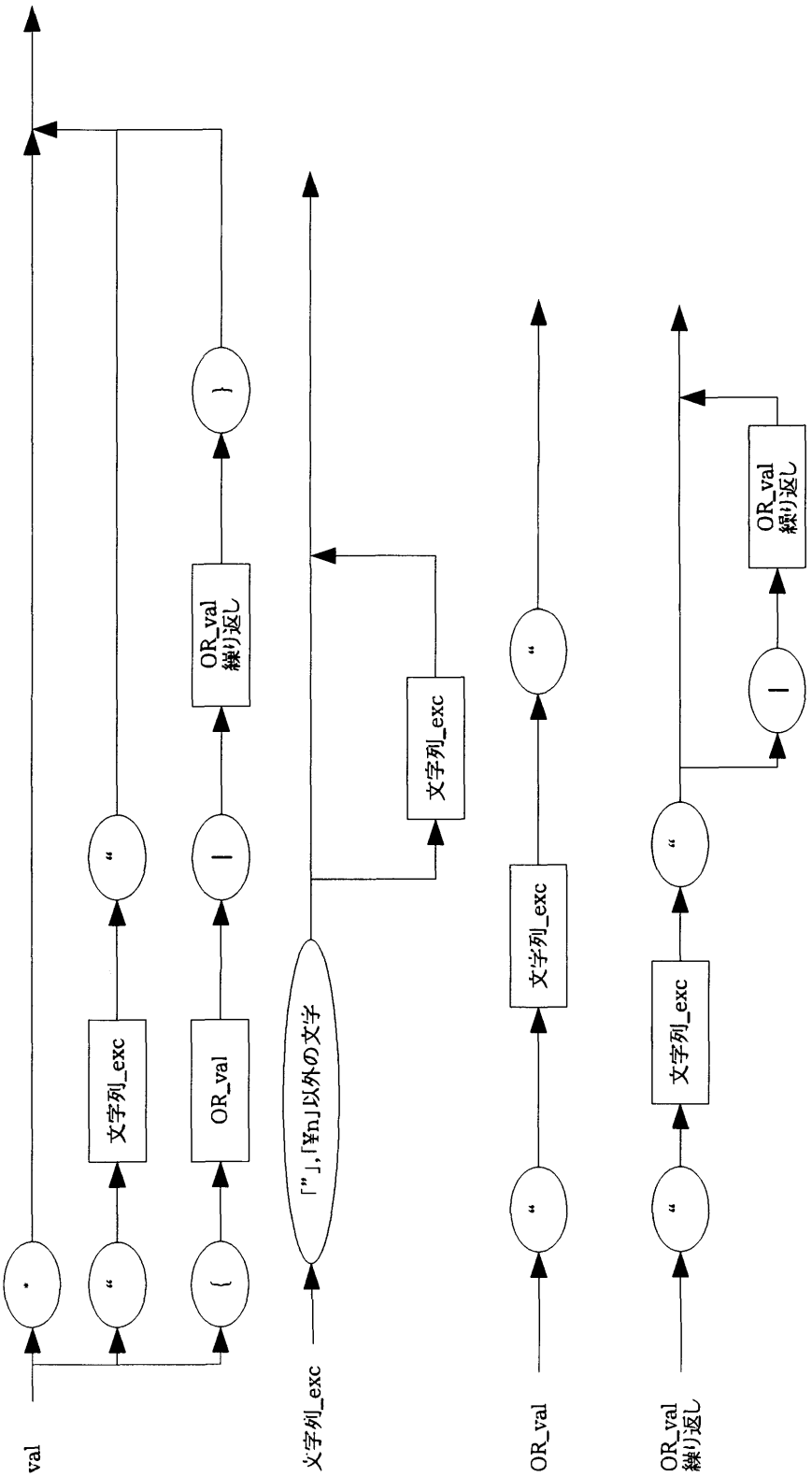
書き換え前: <date abs="19980000">1998 年</date>に<val>100 円</val>まで上昇した

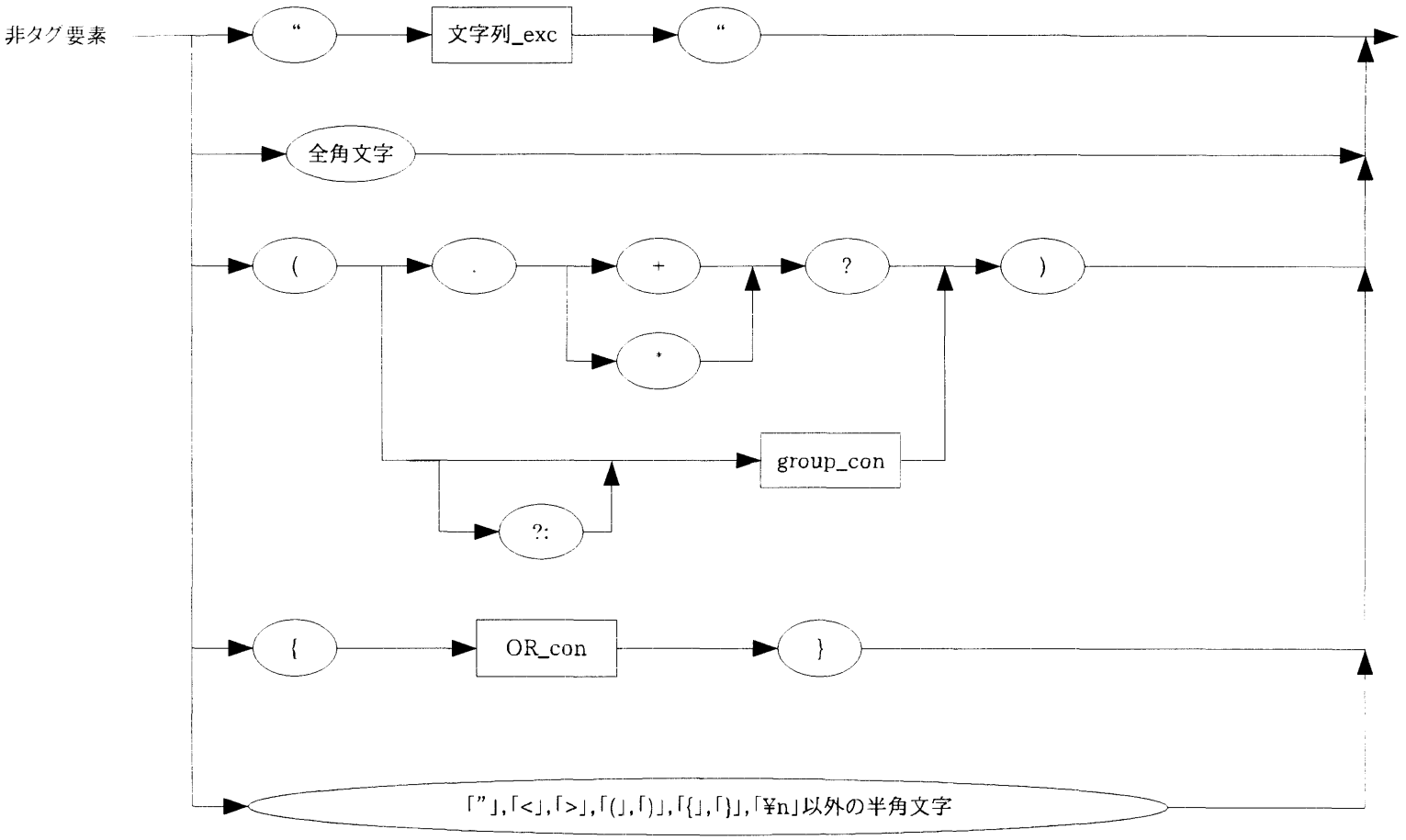
書き換え後: <ipc date=19980000 type0="dur_eq" val="100 円" type1="upward"><date abs="19980000">1998 年</date>に<val>100 円</val>まで上昇した</ipc>

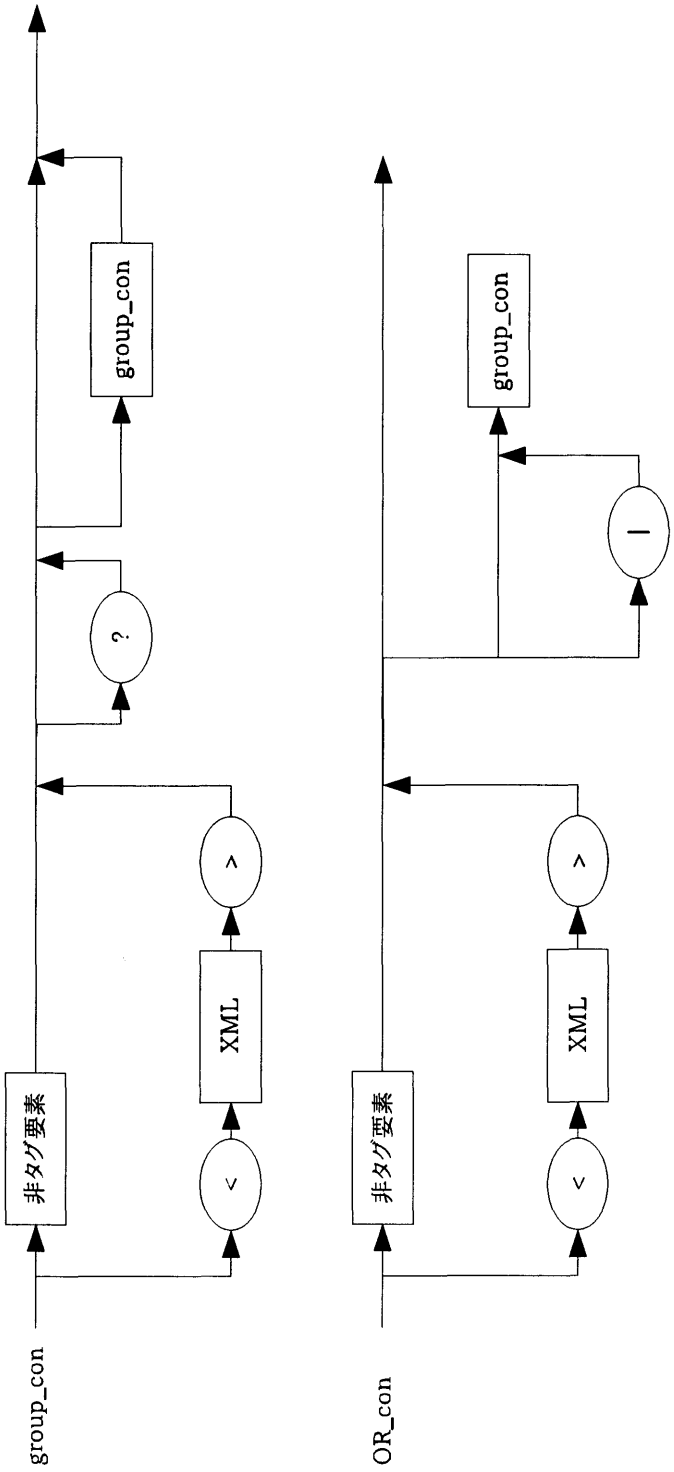
図 B-2　書き換えの例

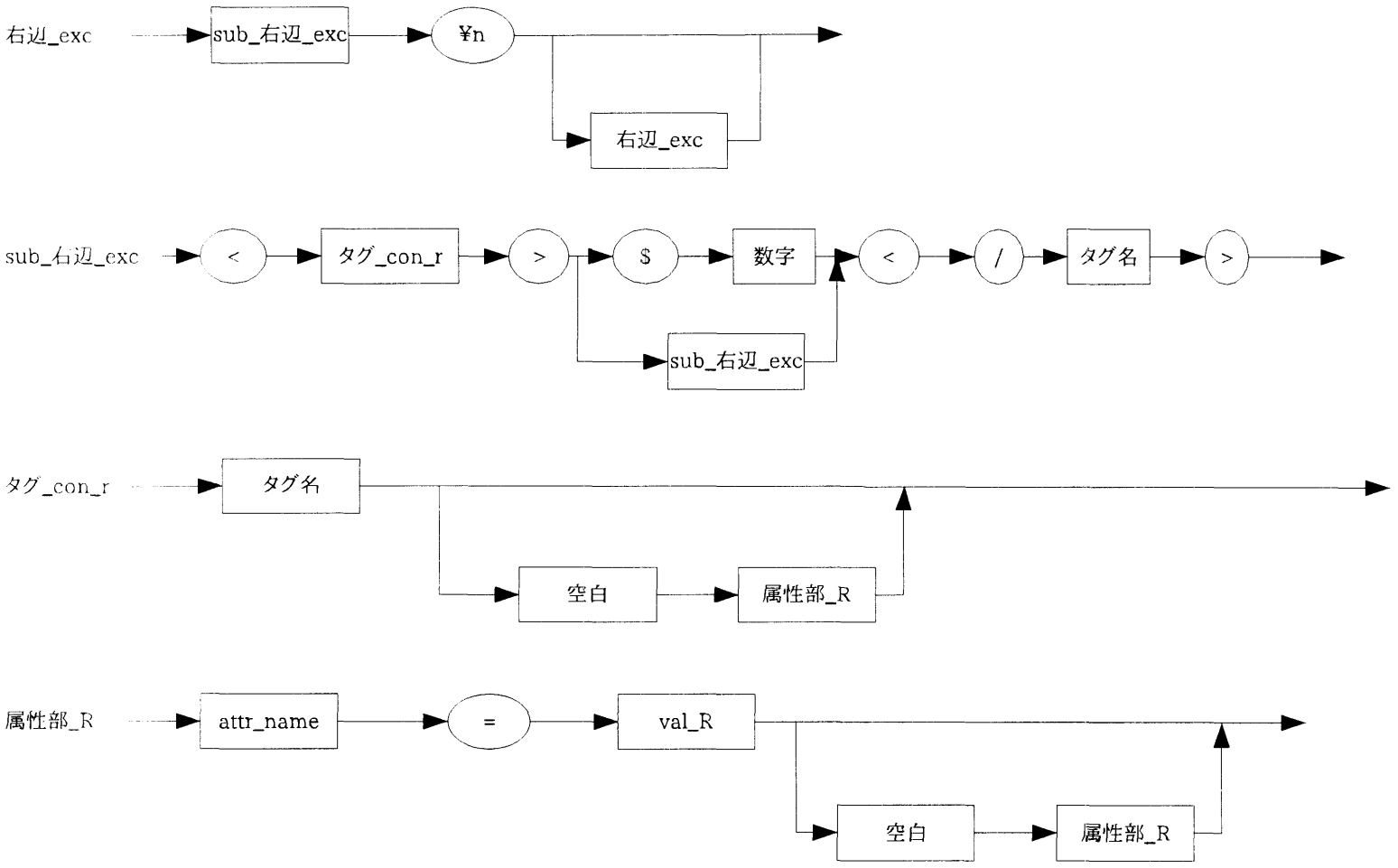

書換え規則
左辺_exc
==>
右辺_exc
左辺_exc
<
¥
タグ_con
XML
>
?
左辺_exc
非タグ要素
XML
<
XML_タグ
>
非タグ要素
<
/
XML_タグ
XML_タグ
<
XML
>
XML
「<」,「>」,「¥u」,「¥」,「/」以外の文字
XML_タグ

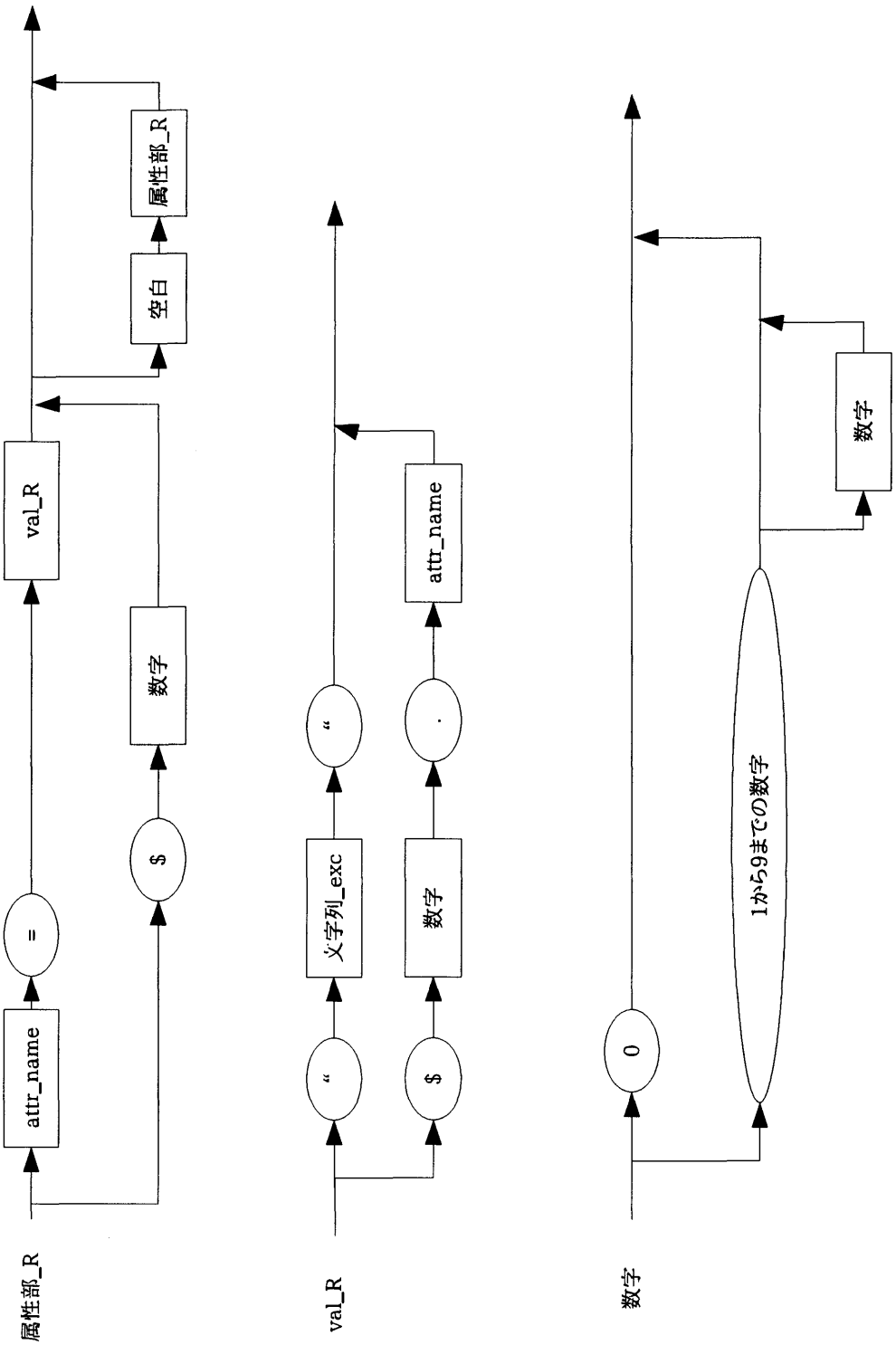

タグ_con
タグ名
空白
属性部_L
属性部_L
attr_name
=
val
!
空白
属性部_L
空白
半角スペース,タブ
空白
タグ名
「<」,「>」,「¥n」,半角スペース,タブ以外の文字
タグ名
attr_name
「=」,「!」,「¥n」,「<」,「>」,半角スペース,タブ以外の文字
attr_name


val
'
"
{
文字列_exc
OR_val
}
OR_val
繰り返し
文字列_exc
「"」,「¥n」以外の文字
文字列_exc
"
文字列_exc
OR_val
"
文字列_exc
"
|
OR_val
繰り返し


非タグ要素
“
文字列_exc
“
全角文字
(
.
+
?
)
*
group_con
?:
{
OR_con
}
「”」,「<」,「>」,「(」,「)」,「{」,「}」,「¥n」以外の半角文字


group_con
非タグ要素
<
XML
>
?
group_con
OR_con
非タグ要素
<
XML
>
|
group_con


右辺_exc
sub_右辺_exc
¥n
右辺_exc
sub_右辺_exc
<
タグ_con_r
>
$
数字
<
/
タグ名
>
sub_右辺_exc
タグ_con_r
タグ名
空白
属性部_R
属性部_R
attr_name
=
val_R
空白
属性部_R


属性部_R
attr_name
=
$
数字
val_R
空白
属性部_R
val_R
“
文字列_exc
“
$
数字
.
attr_name
数字
0
1から9までの数字
数字

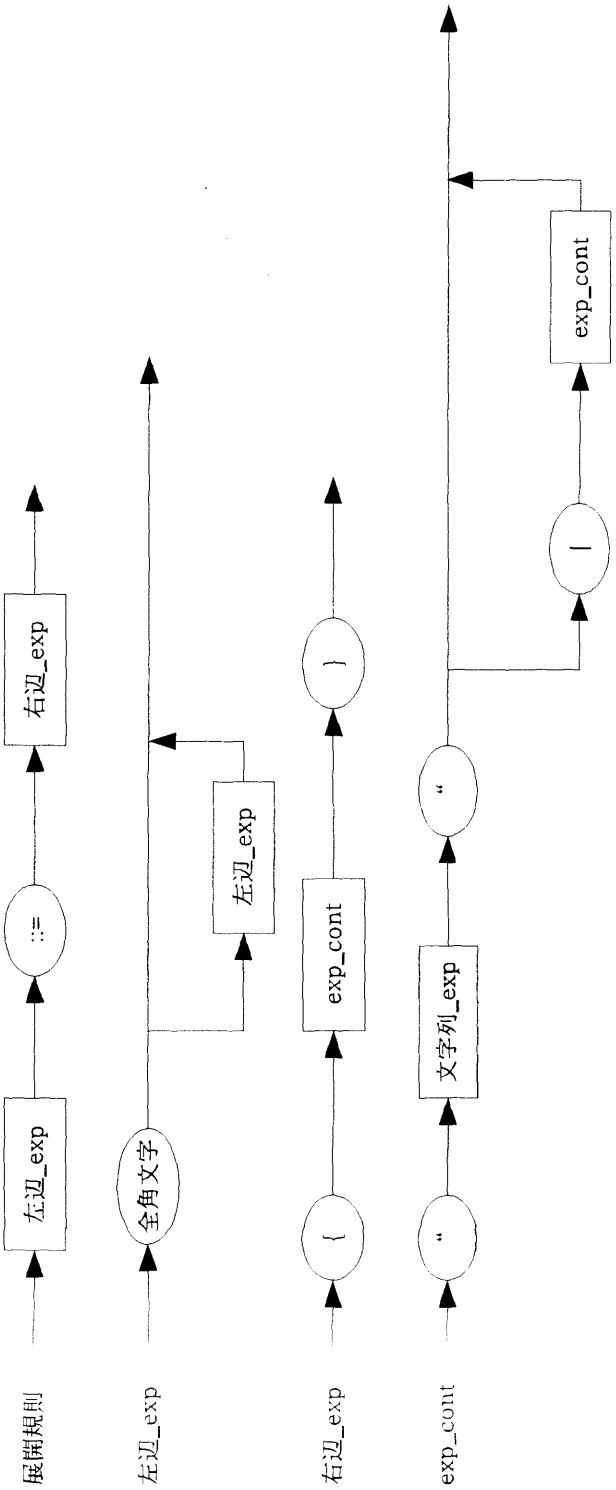

展開規則
左辺_exp
::=
右辺_exp
左辺_exp
全角文字
左辺_exp
右辺_exp
｛
exp_cont
｝
exp_cont
“
文字列_exp
“
|
exp_cont

図 B-3　構文図式