相対表現に基づいた動向情報抽出システムの構築

上西 康広† 松葉 達明‡ 桝井 文人† 河合 敦夫† 井須 尚紀†
†三重大学大学院工学研究科 ‡三重大学工学部
{*uenishi, matsuba, masui, kawai, isu*}*@ai.info.mie-u.ac.jp*

1 はじめに

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 [2] 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 [3]」が注目されている. 動向情報とは「06 年からのゲーム業界はどうなっているか」, 「今年のガソリンの価格はどう動いているだろう」といった, ユーザの関心に対する最初の回答となるものである. 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, 地図などを用いてまとめて視覚化した方が直感的に理解し易い. 以上を踏まえると, テキスト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れば, 動向情報を効率的に把握する支援技術として非常に有効である.

テキスト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が, 我々は最も基本的である stage 1 に取り組む.

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: 注釈などのテキスト情報を用いた可視化情報の補足

以下, 動向情報の要約と可視化に関するワークショップ (MuST) より公開されている MuST コーパス [3] 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.

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り, 難波ら [4] は, これらの表現を**相対表現**と呼んでいる. 図 1 に相対表現の例を示す. 相対表現は, 他に示された情報を参照し, 比較することによって, 相対的に他の情報を示す機能を持つ. 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, 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.

例えば, テキストに明示されている 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). 相対表現「前年比 0.1%増」を利用すれば, 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) を把握することができる.

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 1: 相対表現と 4 つ組の例

Uenishi et al.[5] は相対表現を利用して新聞記事から動向情報を抽出する手法を提案し, システムの有効性を検証している. しかしながら, クエリと name 要素との関連性判定の性能に問題があった.

例えばクエリとして「パソコン国内出荷台数」が与えられた場合, 「国内パソコン出荷台数」とは関連がないと判定される. これは, 「パソコン/国内出荷台数」, 「国内パソコン/出荷台数」のように文字列を単純に二分割し, 前後の各文字列同士の一致を見るだけなので, 同じ意味でも形態素の並びが異なる文字列には対応できないためである.

本論文では, この問題を改善するシステムを提案する. 提案システムは, クエリのバリエーションを増やすことでこの問題に対処する.

以下, 2 章で提案システムの概要について述べ, 3 章で提案システムの評価実験について述べる. 4 章で実験結果についての考察を行う.

2 提案システムの概要

本章では, 提案システムの概要について述べる. 本システムは, 相対表現に基づいて, 入力された統計量名 (クエリ) と関連する explicit な 4 つ組を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. 本システムは, 以下の 4 つのモジュールから構成される.

module 1: クエリ解析・文書検索
module 2: explicit な 4 つ組の抽出
module 3: 4 つ組の選択
module 4: implicit な 4 つ組の生成

以下, 各モジュールの処理について述べる.

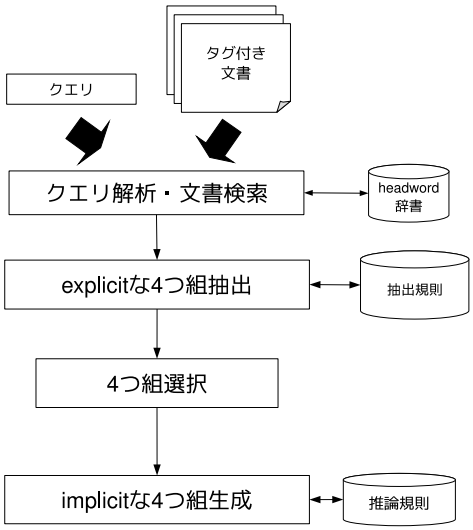

図 2: システム構成

2.1 クエリ解析・文書検索

本モジュールでは, 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 speci er 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 3 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, 各クエリを名詞・未知語・接頭詞を抽出する.

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形態素の系列中の形態素の 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.

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

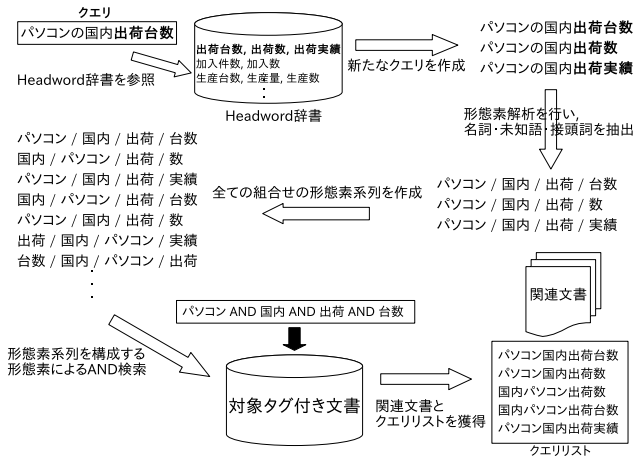

図 3: クエリ解析・文書検索の概要

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

Ex2.
<name> は <date> 比 <rel> 減の <val>

図 4: 抽出規則の例

2.2 explicit な 4 つ組の抽出

本モジュールでは, 抽出規則を用いて, 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規則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.

2.2.1 基本要素抽出

抽出規則を用いて, 4 つ組を構成する基本要素と相対表現を抽出する. 抽出規則は, 今岡ら [1] の方法に基づき 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 に抽出規則の例を示す.

対象文書が入力されると, まず, 文書中に抽出規則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 5 の例では, 対象文書は図 4 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.

2.2.2 不足要素の補完

抽出規則だけでは, 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規則のみで抽出を行い, 補完を行わない.

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。

⇑

抽出規則:
<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 5: 基本要素の抽出

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 6: 不足要素の補完例

name: 抽出規則の適合箇所の前方, かつ, 最近傍の name 要素

par: 抽出規則の適合箇所の前方, 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) 抽出規則の適合箇所の前方, 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 6 の例文において抽出規則「<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 要素は抽出規則では抽出されないため, 補完処理を行う. この場合 par 要素は抽出規則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.

2.3 4 つ組の選択

本モジュールでは, 獲得できた 4 つ組からクエリと関連する 4 つ組を選択する. 選択には 2.1 節のクエリ解析で獲得した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配列 N,P を生成する.

step 2 配列 P を配列 N の先頭に結合し配列 Q を生成する.

step 3 リストからエントリ (要素数 A の配列 q) を取り出し, 配列 Q(要素数 B) と要素の表層情報を用いて関連性を判定する. 判定に用いる条件は以下の通りである.
(1) 配列 q が配列 Q から生成される長さ A の部分配列 Q'_i のいずれかに一致する場合.
(2) 配列 Q が配列 q から生成される長さ B の部分配列 q'_i のいずれかに一致する, かつ, 配列 q の残りの要素が name 要素の抽出文中に存在する場合.

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.

例えば, 配列 Q「NEC/パソコン/国内/出荷/台数」(要素数 5) と配列 q「パソコン/国内/出荷/台数」の場合, 配列 q は配列 Q 中から生成される長さ 5 の部分配列 (パソコン/国内/出荷/台数) と一致するので関連があると判定される.

2.4 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を用いて implicit な 4 つ組の生成する. 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.

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

⇓

{ ビール出荷量, アサヒ, 2007 年,1 億 8824 万ケース }

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.

$$\begin{aligned} date_{imp} &= \text{2007 年 - 1 年} \\ &= \text{2006 年} \\ val_{imp} &= \frac{\text{1 億 8824 万ケース}}{1 + \frac{0.1}{100}} \\ &= \text{1 億 8800 万ケース} \end{aligned}$$

⇓

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

3 評価実験

提案システムの有効性について検証するため従来システム [5] との比較実験を行った.

以下, 実験環境について述べる.

入力データとして, NTCIR MuST T2N subtask[3] で配布された MuST コーパス (XML タグ付き毎日新聞,1998 年〜2001 年,120 記事) を用いた. 同様に, 上記タスクで用いられたクエリ (20 個) を用いた.

テストデータからクエリと関連する 4 つ組が正しく抽出されたかを評価した. 本システムは, 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する. また, 本実験では explicit な 4 つ組のみを評価対象とする. 評価値として適合率, 再現率, F 値を用いた.

表 3 に提案システムと従来システムの有効性評価の結果を示す. 提案システムは適合率 0.556, 再現率 0.338, F 値 0.421 となり, 従来システムと比べて適合率は低下したものの再現率, F 値が向上した. F 値の差について approximate randomization 検定 [6] を行ったところ有意水準 5%で有意差が確認された.

表 1: 有効性の評価

	従来システム	提案システム
適合率	0.638(30/47)	0.556(45/81)
再現率	0.226(30/133)	0.338(45/133)
F 値	0.333	0.421

4 考察

F 値の差に統計的に有意な差が確認されたことから, 提案システムは従来システムよりも有効であることが示唆される. これは, クエリ解析においてクエリのバリエーションが増加したことから, 4 つ組の選択は効果的に働いたと考えられる.

次に失敗例として, 「内閣支持率」というクエリに対して「内閣不支持率」の 4 つ組を選択してしまった例が上げられる. システムはクエリから生成される形態素の配列が比較対象から生成される形態素の部分配列と一致するかを見ている. この場合, headword が異なるにもかからず「内閣支持率」は「内閣不支持率」の部分配列と一致し, 関連があると判定されたことが原因である. これに対応するためには, 配列の一致の判定を headword と speci er で個別に実施する必要がある.

システムの性能向上のためには, 本論文では改良を加えなかった emplicit な 4 つ組の抽出の性能を向上させる必要もある. 4 つ組の抽出は基本的には相対表現の出現パターンを拡張させた抽出規則によるものである. しかし, 「ガソリン」トピックのような, 相対表現の出現の仕方が発散しており, パターン化が困難であるトピックに対しては, 異なるアプローチが必要である.

5 おわりに

本論文では, 相対表現に基づく動向情報抽出システムを提案した. 従来システムに対して 4 つ組の選択における関連性判定の改善を行った. 具体的には, クエリのバリエーションを増やすことで 4 つ組選択の性能向上を試みた. 評価実験によって提案システムが従来システムより F 値が向上し, 動向情報抽出における有効性を確認した.

今後は, explicit な 4 つ組抽出の性能向上に取り組んでいきたい. また, タグ付き文書ではなくプレーンな文書にも対応できる手法を考えていきたい.

謝辞

本研究は科研費（20500833）の助成を受けたものである。

参考文献

[1] 今岡裕貴, 桝井文人, 河合敦夫, 井須尚紀. 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, Vol. 18, No.5, pp.735-744, 2006.

[2] 加藤恒昭, 松下光範. 情報編纂（Information Compilation）の基盤技術, 第 20 回人工知能学会全国大会, 1D3-2, 2006.

[3] Kato,T., and Matsushita,M. Overview of MuST at the NTCIR-7 Workshop –Challenges to Multimodal Summarization for Trend Information–, *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 475-488, 2008.

[4] 難波英嗣, 国政美伸, 福島志穂, 相沢輝昭, 奥村学. 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, NL-168, pp.67-74, 2005.

[5] Uenishi,Y., Masui,F., Matsuba,T., Kawai,A., and Isu,N. Trend Information Extraction based on Relative Expression participated on MuST T2N Subtask, *Proceedings of the 7th NTCIR Workshop Meeting*, pp. 509-514, 2008.

[6] Chinchor,N., Lewis,D.D., and Hirschman,L. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3), *Computational Linguistics*, 19(3): pp.409-449, 1993.