複数文書からの動向基本情報抽出における相対表現の有効性の検討

今岡裕貴　桝井文人　河合敦夫　井須尚紀
三重大学工学部情報工学科　　〒514-8507　三重県津市栗真町屋町 1577
E-mail: {imaoka,masui,kawai,isu}@ai.info.mie-u.ac.jp

1. はじめに

「動向情報の可視化」技術が注目されている[3]. これは, 電子化された大量の情報源に対して, ユーザが欲する特定情報に効率良くアクセスするための支援技術である.「動向情報」とは, ある商品の売り上げ, ある業界のメーカー別業績, 内閣の支持率など, 時々刻々と変化する数値に関連した情報である. したがって, 動向情報は, 文章で説明するよりも, グラフや図, 地図などを用いて視覚化した方が直感的に理解し易いケースが多い. そのため, テキスト中に記述された動向情報を効率的に把握する支援技術として, テキスト情報を解析して動向情報を自動抽出し, さらに, ユーザの目的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. 最近では, このような技術研究の発展・確立を目指した『動向情報の要約と可視化に関するワークショップ“MuST”(Workshop on Multimodal Summarization for Trend Information)』[3]が開催され, 同技術研究の加速が期待されている.

1. **ビール大手 5 社が 12 日まとめた 1997 年の課税出荷数量は 5 億 3379 万ケースと前年比 97.8%にとどまった.**
2. **1999 年 1 月のビール全体の出荷量は 3056 万ケースで, 前年同月比 12.5%増となった.**

図 1：新聞記事における相対表現の例

このような動向情報の一部として現れる数量表現や時間表現には, 数値の相対的な差異や, 数値の変動として表現されるものがある. 図 1 にそれらの例を示す(下線部). 難波ら[2]は, これらの表現を相対表現と呼び, これらの表現を利用することによって, 間接的により多くの動向情報を収集することができると指摘している. 今岡ら[5][6]は, 新聞記事中に現れる相対表現を利用して動向に関する基本情報を抽出する規則を構築し, 実験によってその有効性を明らかにしている.

動向情報の収集効率を向上させることを目的とした場合, 複数文書を利用して抽出規則を構築する方法が考えられる. 難波ら[2]は, 複数文書からの動向情報抽出の有効性も検証している. 彼らは, 動向情報の抽出を一種の複数文書要約と考え, あるトピックに関する複数文書から動向情報の自動抽出を試みている.

動向情報抽出が一種の複数文書要約であるという考えに立脚するなら, 相対表現を利用した基本情報抽出も, 複数文書を利用した方が有効であるという仮説を考えることができる. しかしながら, 難波らの研究では, 相対表現は扱われていない. また, 今岡ら[6]の研究でも, 単一文書からの抽出規則が対象となっており, 複数文書を対象とした場合の有効性につては論じられていない.

そこで本論文では, 複数文書からの動向基本情報の抽出に関して, 相対表現を用いた場合の効果について議論する. 相対表現を利用して構築した基本情報抽出規則を対象とし, (1)ある新聞記事コーパスから構築した抽出規則を, 同紙の新聞記事コーパスに適用した場合と異紙の新聞記事コーパスに適用した場合に抽出性能の差が見られるか, (2)単一の新聞記事コーパスから構築した抽出規則を適用した場合と複数の新聞記事コーパスから構築した抽出規則を適用した場合に抽出性能の差がみられるか, について分析する. (1)(2)の結果より, 複数文書からの動向基本情報抽出における相対表現の有効性について論じる.

以下, 2 章で動向情報の要素となる基本情報, および相対表現の定義について述べ, 3 章で相対表現を利用した基本情報抽出方法について説明する. 4 章で基本情報抽出手法を複数文書に適用した場合の評価実験を行い, 5 章でその有効性について考察する.

2. 動向情報と相対表現

本章では, 動向情報の「基本情報」を定義する. さらに, 相対表現中で基本情報が出現する形態, および基本情報相互の関連について説明する.

2.1. 基本情報と相対表現

テキスト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. 我々は, MuSTワークショップにおいて公開されているMuSTコーパスの注釈仕様[4]を参考にして, 3 種類の要素,「統計量名(LABEL)」,「数値情報(NUM)」,「時間情報(TIME)」を「**基本情報**」として定義する. 本論文では, これらの要素をまとめて 3 つ組と呼び, (LABEL, TIME, NUM) のように記述することにする.

基本情報は, タグによって明示的に示されていない場合がある. そのような場合の典型例として相対表現があげられる. 図 1 に相対表現の例を示す. 相対表現では, 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. したがって, 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数値情報を推論することができるはずである.

例えば, 図 1 の例文 1 において, 基本情報の 3 つ組 (LABEL, TIME, NUM)として,(ビールの課税出荷数量, 1997 年, 5 億 3379 万ケース) の関係が把握できているとする (図 2). その場合,「前年比 97.8%」という相対表現からは,

時間情報(TIME)：前年 = 1997 年 － 1 年
= 1996 年

数値情報(NUM)：97.8％ = 5 億 3379 万ケース/0.978
= 5 億 4580 万ケース

という推論が可能である．その結果，新たに（ビールの出荷数量，1996 年，5 億 4580 万ケース）という基本情報の 3 つ組関係が把握できる（図 3)．

LABEL	：（ビール大手 5 社の）課税出荷数量
TIME	：1997 年
NUM	：5 億 3397 万ケース

図 2：図 1 から抽出できる基本情報の 3 つ組

LABEL	：（ビール大手 5 社の）課税出荷数量
TIME	：1996 年
NUM	：5 億 4580 万ケース

図 3：相対表現を利用して得られる基本情報

<date>1999 年 1 月</date>の<name>ビール全体の出荷量</name>は<date>前年同月</date>比<rel>12.5％</rel>増の<val>3056 万ケース</val>となった

↓

<TIME1> の **<LABEL>** は **<TIME2>** 比 **<REL>** 増 の **<NUM>**となった

図 4：出現パターンの抽象化の例

3. 相対表現を利用した基本情報抽出

本章では, 相対表現を利用した複数文書からの基本情報の抽出手法について説明する.

3.1. 抽出規則の獲得と基本情報の抽出

訓練コーパスから獲得した出現パターンの抽象化の例を図 4 に示す. 図のような出現パターンが得られたとすると，その中に記述された基本情報部分を抽象化する．抽象化された出現パターン（抽象パターン）には類似したものも存在する．そこで，類似した抽象パターン相互に形式的な階層関係を持たせる. 抽象パターンを階層化した例を図 5 に示す．この場合,「<TIME>比<REL>」に関連する抽象パターンが関連付けられている．階層化により，関連する抽象パターンの優先順位が決まる．

上で述べた抽出規則を用いて動向をあらわす基本情報を抽出する過程を説明する. 処理対象のテキストが入力されると，まず，テキスト中の文字列を捜査して，抽出規則に合致する箇所が存在するかどうかを調べる. 規則に合致する箇所が発見された場合, 抽象化パターンとの照合を行う. 照合に成功した場合は基本情報を抽出する. この結果, 動向に関する基本情報の 3 つ組(LABEL, TIME, NUM)が得られる. 図 1 の例文 2 が抽象化パターン“ <TIME>の<LABEL>は<TIME>比<REL>{増|減}<NUM> ”に照合した場合を考えると，このとき，まず 3 つ組として,
（ビール全体の出荷量，1999 年 1 月，3056 万ケース）
が得られる.

抽象化パターンと基本情報が決定した後, 間接情報の推論規則を用いて, 間接的に示されている基本情報を推論する．推論規則は，抽象化パターンに関連付けられている. 推論規則には, フローの条件に従って辿る処理経路に応じて異なる計算規則が対応する. 前述の例に対応して適用される間接情報の推論規則の例を図 6 に示す. この例の場合, (1)が適用され,

3056 万ケース / （1 ＋ 12.5 / 100）= 2716 万ケース

が推論され，最終的に
（ビール全体の出荷量，1998 年 1 月，2716 万ケース）
という 3 つ組が得られる.

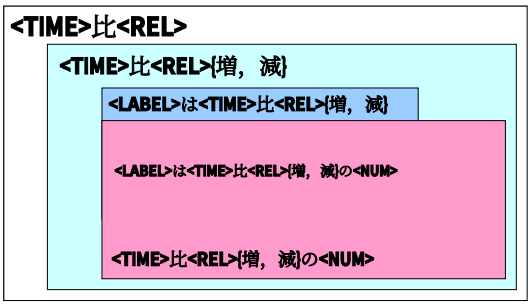
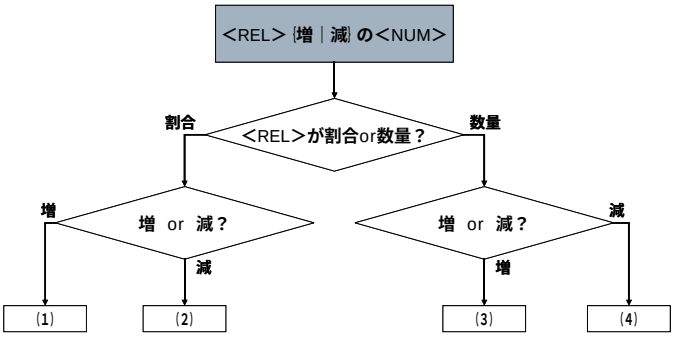


図 5：抽象パターン階層化の例



（1）NUM	／ （1＋（REL／100))
（2）NUM	／ （1－（REL／100))
（3）NUM	－ REL
（4）NUM	＋ REL

図 6：間接情報の推論規則の例

3.2. 複数文書を用いた抽出規則構築

上記のような手順で複数文書から抽出規則の構築を考えた場合，抽出規則間で共通する規則と，共通しない規則が存在することが予想される. その場合, 共通する規則は, 文書を横断して適用可能な, より汎用性の高い抽出規則である可能性が高い．一方，共通しない規則は，文書に依存している規則である可能性が高い.

さらに演繹すると，複数文書を利用した結果，共通規則が多く得られることが確認されれば，複数文書の利用は，汎用規則獲得において有効ということになる．反対に，共通する規則がほとんどない場合は，複数文書の利用は，文書に依存した抽出規則獲得に有効であることになる. しかし, 複数文書利用がいずれの有効性を持つものであるかを確かめるためには，実際に分析評価を行う必要がある.

4. 実験と評価

前章で述べた基本情報抽出手法が，複数文書を対象にした場合の有効性について検証するための評価実験を行った．

対象とするコーパスとして，1998 年，1999 年の毎日新聞（168 記事）と読売新聞（265 記事）を用いた[1]．これらのコーパスから，それぞれ人手で出現パターンを取り出し，それらを抽象化し，抽出規則を作成した．その結果，毎日新聞から 93 組（以下，毎日規則），読売新聞から 124 組（以下，読売規則），両コーパスから 179 組（以下，混合規則）の抽出規則が得られた．

表 1 に各トピックに対して，新聞 2 紙の記事数とそれぞれの新聞から得られた規則数, 共通する規則数を示す．毎日新聞から得られた規則数が一番多いトピックは「住宅(Housing)」の 37 組，一番少ないトピックが「株価（Nikkei)」の 9 組であった．また，読売新聞から得られた規則数が一番多いトピックは「株価（Nikkei)」の 43 組，一番少ないトピックは「支持率（Politics)」の 4 組であった．共通規則数が一番多いトピックは「ビール（Beer)」「パソコン(PC)」「失業率（Unemploy)」の 13 組，一番少ないトピックは「支持率（Politics)」の 1 組であった．

構築した抽出規則を，毎日新聞，読売新聞に適用して基本情報を抽出し，その抽出性能を評価・比較した．トピック毎に得られた基本情報がそれぞれ正しく獲得できたかどうかに基づいて，F 値を算出した．毎日新聞と読売新聞に各規則を適用したときのマクロ平均 F 値とマイクロ平均 F 値を表 2 に示す．毎日新聞で最も値が高かった F 値をあげると，マクロ平均 F 値，マイクロ平均 F 値ともに毎日規則の 0.828, 0.842 であった. 読売新聞では，マクロ平均 F 値，マイクロ平均 F 値ともに混合規則の 0.716，0.702 であった．

また, 毎日新聞と読売新聞のトピック別の F 値を図 7, 8 に示す．縦軸が F 値で，横軸が各トピックであり，各規則に対してそれぞれの F 値を示している．毎日新聞をみると（図 7)，全体的に毎日規則が高い F 値を示した．また，読売規則を「株価」トピックに対して適用したときの F 値の減少が目立った．読売新聞をみると（図 8)，全体的に読売規則と混合規則が高い F 値を示しており，毎日新聞でみられたほどの差はみられなかった．トピック別でも, 目立った F 値の上昇や減少はみられなかった．

表 1：各コーパスの記事数と得られた抽出規則数

	毎日新聞		読売新聞	
	マクロ平均F値	マイクロ平均F値	マクロ平均F値	マイクロ平均F値
毎日規則	0.828	0.842	0.632	0.621
読売規則	0.640	0.678	0.698	0.682
混合規則	0.765	0.784	0.716	0.702

表 2：各コーパスに対する F 値

	毎日_記事数	読売_記事数	毎日_規則数	読売_規則数	共通規則数
Beer	22	30	25	20	13
Car	16	35	14	41	5
Housing	35	31	37	25	7
Nikkei	37	65	9	43	5
PC	20	21	27	19	13
Politics	17	40	10	4	1
Unemploy	21	43	29	22	13
All	168	265	93	124	38

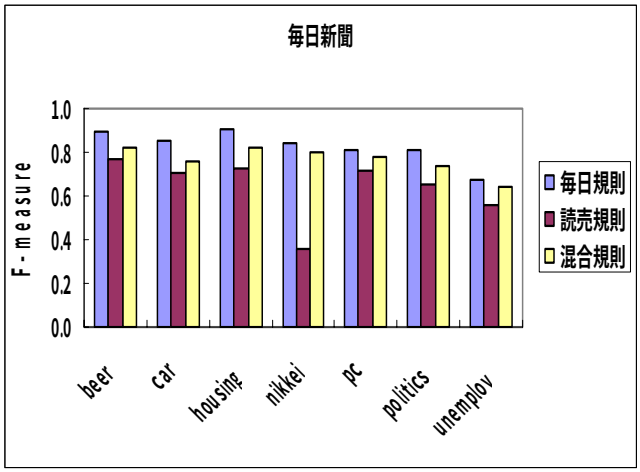


図 7：毎日新聞に対する各規則の F 値

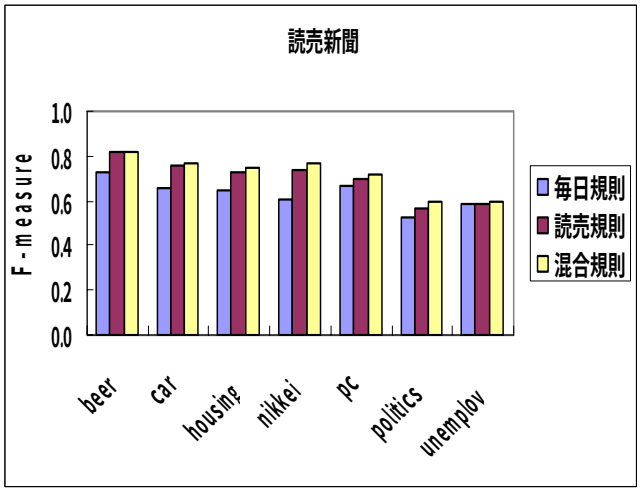


図 8：読売新聞に対する各規則の F 値

5. 考察

実験結果についての考察を行う．最初に，単一の新聞記事コーパスから構築した抽出規則の適用結果についての考察, 次に複数の新聞記事から構築した抽出規則の適用結果についての考察, 最後に特徴的なトピックについての考察を行う．

まず，毎日規則と読売規則の性能について考察する．

毎日規則を毎日新聞記事に適用した場合, トピックを区別せずに抽出性能を F 値で表したマイクロ平均 F 値は 0.842 であり，トピック毎の抽出性能を平均したマクロ平均 F 値は 0.828 であった．二つの数値の差が 0.014（F 値に対して 1%程度）と小さいことから，実験に用いた毎日規則は，全体として偏りなく機能したと考えてよい．

毎日規則を読売新聞記事に適用した場合, マイクロ平均 F 値は 0.632，マクロ平均 F 値は 0.621，二つの数値の差は 0.038（F 値に対して 1%程度）であった．同一紙への適用

[1] 対象トピックは，出現パターンが 10 以上得られた 7 トピックとした．

結果と比較して, 抽出規則自体は偏りなく機能したと考えられるが，F 値が 0.1 ポイント以上低い値であり，毎日規則では網羅できない相対表現が読売新聞記事に存在していることがわかる．

同様に, 読売規則を用いて同一紙に適用した場合, 毎日新聞記事に適用した場合の結果をみると, 毎日規則の結果と比較して, 同一紙に対する抽出性能は 0.1 ポイント以上低い結果となった．しかし，他紙（毎日新聞記事）に対する抽出性能は大きく変わらなかった．

以上より，毎日新聞記事では, 比較的一般性の高い相対表現が多用され, 読売新聞記事では相対表現のバリエーションが多く，トピック依存性が強いといえる．さらに，単一新聞記事のみを用いて抽出規則を構築した場合, 基本的な抽出性能を確保することは可能であるが, より網羅性の高い抽出規則を得ることについては限界が生じる．よって, 相対表現を介した場合に置いても, 一般性の高い抽出規則を得るには, 複数新聞記事を用いた方がよいと結論できる．

次に, 混合規則を適用した場合について述べる．混合規則の適用結果は，毎日新聞記事に対しては，マイクロ平均 F 値が 0.784，マクロ平均 F 値が 0.765 であった．読売新聞記事に対しては，マイクロ平均 F 値が 0.702，マクロ平均 F 値が 0.716 であった．これらの場合も，マイクロ平均とマクロ平均の差は F 値に対して 1.5〜2%程度であったので，抽出規則はトピックに関係なく機能したといえる．また，混合規則を読売新聞記事に適用した結果では，同紙単一記事に基づく抽出規則の性能を F 値で上回った. これは, 単一記事では網羅しきれなかった抽出規則が, 複数記事を用いることによって補完された効果である．

二紙から得た二つの抽出規則の共通部分のみを取り出した抽出規則（共通規則）についての性能についても吟味する．共通規則は 38 組得られ，毎日規則の約 40%, 読売規則の約 31%を占めている．共通規則を毎日新聞記事に適用した結果をみると，マイクロ平均 F 値が 0.549(precision 0.702, recall 0.451)，マクロ平均 F 値が 0.537(precision 0.681, recall 0.476)であった．読売新聞記事に適用した結果では, マイクロ平均 F.値は 0.427(precision 0.656, recall 0.316), マクロ平均 F 値が 0.416(precision 0.678, recall 0.313)であった．いずれの結果についても precision については 0.656 から 0.702 の性能が得られており, 基本的な抽出性能は得られたとみてよい．

最後に，特徴的なトピックについて考察する．読売新聞記事の「支持率」トピックから得られる規則は記事あたり 0.1 と少なかった．これは，図 9 に示すように読売新聞記事の「支持率」トピックでは，動向情報が相対表現として表現されない傾向が強いためである．

図 7 をみると，「株価」トピックにおいて，読売規則を毎日新聞記事に適用した場合の性能が目立って低い. これは, 読売規則が毎日新聞記事の相対表現に通用しなかったことを意味する．一方で，毎日規則を読売新聞記事に適用した場合の性能に大きな落ち込みは見られない．これは，毎日規則が読売新聞記事の相対表現に対して通用したことを意味する．

総合的には，毎日規則は相対的にみて汎用性が高く，読売規則は汎用性が低い性質がみられるため，「株価」トピックについてもこの傾向が強く現れたようにみえる. しかし，両紙の「株価」トピック記事について調べると，毎日新聞記事では, トピック固有の相対表現が用いられており, 読売新聞記事では, トピックに依存しない一般的な相対表現が用いられる傾向にある（図 9）．したがって，実際には，毎日規則の「株価」固有規則は機能せず，一般性の高い規則が適用された結果であることがわかる．

以上のことから，相対表現を利用した動向基本情報抽出において，単一文書から抽出規則を構築するより，複数文書から抽出規則を構築する方が有効であることが確認された．また，複数文書を用いた場合，得られる共通規則は，他の文書においても適用されることが多い汎用的な規則であると結論できる．したがって，全ての文書に適用する共通規則と，それぞれの文書の固有規則を用意し，適切に使い分ければ，より安定した動向基本情報抽出が可能となる．

「支持率」
毎日新聞：先月調査比 5 ポイント減の 10%
読売新聞：今回の調査は 10%（前回 20%）

「株価」
毎日新聞：前日終値比 100 円安の 1 万 4000 円
読売新聞：終値は前日比 100 円安の 1 万 4000 円

図 9: 新聞における表記の差異

6.まとめ

本論文では，複数文書からの動向基本情報の抽出に関して, 相対表現を用いた場合の有効性について検証した．ある新聞記事コーパスから構築した抽出規則を，同紙の新聞記事コーパスに適用した場合と異紙の新聞記事コーパスに適用した場合に抽出性能の差が見られるかについて分析した．また，単一の新聞記事コーパスから構築した抽出規則を適用した場合と複数の新聞記事コーパスから構築した抽出規則を適用した場合に抽出性能の差がみられるかについて分析した．　その結果，単一文書から抽出規則を構築する場合より，複数文書から抽出規則を構築した場合が動向基本情報抽出の性能を安定させることが可能であると確認され，複数文書からの相対表現を用いた動向基本情報抽出が有効となりえることを示した．

今後は，対象を相対表現以外にも広げ，間接的に動向情報を収集できる手法の拡張を行って行く予定である．

参　考　文　献

[1] 松下光範，加藤恒昭, “動向情報に基づく情報可視化の基礎検討，” 人工知能学会全国大会，2005.

[2] 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

[3] 加藤恒昭，松下光範，平尾努, “ 動向情報の要約と可視化に関するワークショップの提案, ” 情報処理学会研究報告, vol.2004，no.108(2004-NL-164) pp.89-94，2004

[4] MuST 注釈:
http://www.kecl.ntt.co.jp/scl/workshop/must/spec.html

[5] 今岡裕貴，桝井文人，河合敦夫，井須尚紀: “ 相対表現からの統計情報の導出と提示 ”，電子情報通信学会，NLC2005-120, pp. 37-42 (2006).

[6] 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).