パラレルコーパスを用いた動向情報抽出システムの構築

鈴木 宏哉　斎藤 博昭
慶應義塾大学大学院 理工学研究科
Email: {susuki, hxs}@nak.ics.keio.ac.jp

1　はじめに

近年，新聞記事のような時系列の文書から，一連の事象に関する情報を抽出し，その推移を要約や可視化によって提示する研究に関心が集まっている [1]．この一連の事象に関する情報は動向情報と呼ばれている．

動向情報には,「内閣支持率」や「ガソリン価格」のように時間の経過と共に数値が変動するものと，地震や台風のように，前述の時間と数値に加え，その事象の発生した地点の空間的情報を含むものがある．本稿では，後者の空間的情報を含む動向情報，特に地震情報を対象とした．

また，動向情報は，ある時点での「内閣支持率」のように時間と数値だけで一意に決定されるものばかりではなく，その他の副次的な情報が含まれるものもある．地震に関する動向情報の例を挙げると,「震度」だけでなく,「マグニチュード」のような数値情報も同時に公開される．更に，その地震によって引き起こされた余震の回数なども副次的な情報として得ることができる．この副次的な情報は多岐に亘るため，単一の情報源から全ての情報を得る事は不可能である．そのため，パラレルコーパスを用いて複数の情報源から情報の補完を行う事が有効と考えられる．更に，パラレルコーパスの特徴を利用する事で，動向情報抽出の中間処理である関連文書検索の精度向上を図る事が可能である．

そこで，本稿では 2 紙の新聞記事から地震に関する動向情報を抽出し，提示するシステムを提案する．

2　動向情報

松下らは動向情報を「ある商品の価格や売上高，ある会社の業績，内閣や政党の支持率の推移など，いくつかの統計量に関する時系列データを基にして，その変化を通時的に捉えて纏め上げるものである」としている [2].

動向情報には，大きく分けて以下の二つ内容があり，いずれか，または両方に関する記述が含まれる．ここでの記述とは，単に数値情報だけでなく，その数値の増加や減少に関する説明も含むものである．

- 統計量のある時点での値に関する記述
- ある事象の発生や推移に関する記述

動向情報の要約と可視化に関するワークショップ*で提供されている研究データセット [3] からそれぞれの例を示す.

統計量のある時点での値に関する記述には，企業の業績動向などがあり,

> ソニーは 7 日の 1998 年 3 月期決算発表で，関連会社などを含む連結段階の経常利益が 4537 億 4900 万円 (前期比 45．2 %増) に達し.

のような，ある時点でのある統計量の値の記述と，加えてその値と過去の値との比較や評価からなっている．そのため，この動向情報の提示にはグラフを用いた可視化などが有効である．

二つ目のある事象の発生や推移に関する記述には，ある年の台風の動向などがあり,

> 中型で並の強さの台風 10 号は，17 日午後 4 時ごろ鹿児島県枕崎市に上陸

のような，個々の出来事に関する記述からなる．この記述の場合，その年に幾つの台風がどこに上陸したかという統計的な記述は中心的な要素ではない．

本システムで扱う地震に関する動向情報は，上記の両方の情報からなる．

3　提案システムの概要

本章では，提案する動向情報抽出システムについて述べる．次節で本システム構成の概要を示し，以降で各処理の詳細について説明する．なお，提案システムは昨年度の動向情報の可視化と要約ワークショップにて発表したシステムを元にしている [4].

昨年度のシステムは，研究データセット中の 13 トピックを対象とし，毎日新聞の 97 年版から 99 年版までの 3 年分の記事を使用して，統計量に関する動向情報の抽出とグラフと要約文の提示を行うものであった．

システムの評価結果は以下のようになっている．

表 1: 関連文書抽出の評価結果

トピック	記事に対する再現率	統計量に対する再現率
総合	0.903 (513/568)	0.951 (328/345)

表 2: 統計量と時間情報の組合せの抽出結果

各結果	適合率	再現率	F 値
(1)	0.760 (213/281)	0.655 (213/325)	0.703
(2)	0.776 (218/281)	0.671 (218/325)	0.720

表 3: 比較表現の抽出結果

適合率	再現率	F 値
0.787 (155/197)	0.583 (155/266)	0.670

3.1 システム構成

提案システムの構成は図 1 のようになる．昨年度のシステムとの最大の相違点は，関連文書抽出にパラレルコーパスを処理する部分が追加された点と，数値情報抽出の際に震度と地名の対応付けを行うための処理が追加された点の 2 点である．しかし，システム全体の処理の流れには変わりはない．従って，最初にクエリを入力し，そのクエリを基に関連文書抽出を行い，得られた関連文書中から適切な数値情報の抽出と提示を行う．抽出された動向情報のうちグラフ化が可能なものについてはグラフによる可視化を行う．

3.2 クエリ解析と文書検索，重要語抽出

まず，ユーザはキーワードと抽出したい動向情報のいずれか，または，その両方をシステムに入力する．キーワードと抽出したい動向情報について前述の例を元に説明する．キーワードとは「神津島」のような動向情報を直接指さない語の事であり，新聞記事中のある事象に限定可能な語の事である．抽出したい動向情報とは,「震度」「マグニチュード」のような統計量を伴うものである．

また，新聞記事には様々な単位の数値情報が含まれており，単一の記事に一つの数値情報だけが含まれる場合は稀である．そのため，求める動向情報に合った統計量の単位を選択する必要がある．今回は地震に対象を絞っているため，統計量と単位の組み合わせを辞書として，事前にシステムに与えている．また，入力時のクエリに単位を含める事で，より確実な検索を可能としている．

これらのクエリを元にシステムは類似度の高い記事を検索する．

得られた記事から，次の関連記事検索のベクトル空間法で用いる重要語の抽出を行う．重要語の抽出については昨年度のシステムと同様である．

3.3 パラレルコーパスを用いた関連文書抽出

関連文書抽出処理では，クエリ解析により抽出した重要語を元に，コサイン類似度を用いたベクトル空間法による関連文書の抽出を行う．

*http://www.kecl.ntt.co.jp/scl/workshop/must/

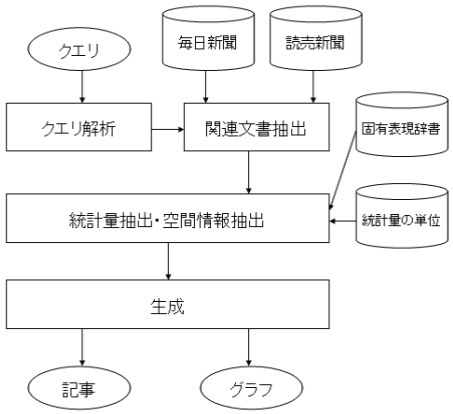

図 1: 提案システムの構成図

まず，パラレルコーパスの対応付けを行い，重要語から作成するベクトルの拡張と，ベクトルの重み付けを行う．

パラレルコーパスとは，同じ内容で表現が異なる文章が対応付けられているコーパスであり，主に対訳コーパスなどを指す．

新聞記事コーパスの場合，社説のように各新聞社の主張が書かれた記事や読者の投稿記事などを除いて，客観的事実を記述した記事については異なる新聞であっても以下のように類似した内容になるという特徴がある．

> 十六日午後七時十二分ごろ，伊豆諸島の神津島で震度３（弱震）の地震があった．気象庁の観測によると，震源地は三宅島近海で深さ一〇キロ，地震の規模はマグニチュード（M）３・２と推定される．
>
> 【毎日新聞 95.1.17】

> 十六日午後七時十二分ごろ，関東沖で地震があり，神津島で震度３（弱震）を記録した．気象庁によると震源地は三宅島近海で，震源の深さは十キロ，マグニチュードは３・２と推定される．
>
> 【読売新聞 95.1.17】

そこで，類似度の高い記事の対応付けを行う事で新聞記事もパラレルコーパスとして扱う事が可能である．

このように複数紙の新聞記事をパラレルコーパスとして用いる事で，関連記事検索の精度の向上を図る事ができる．

これは，ベクトル空間法による関連記事検索の問題点として挙げられる 1 記事に含まれる単語数の少なさによるデータスパースネスの問題を解決し，更にこのパラレルコーパスの両方で出現した単語の重みを上昇させる事で，より類似度の高い記事だけを検索可能になるためである．

実際に，例に挙げた神津島での地震の記事は毎日新聞，読売新聞共に 95 年 1 月 17 日に 1 記事ずつしかなく，それぞれ「伊豆諸島」「関東沖」が一方にしか出現していない．このように，パラレルコーパスを用いる事で重要語ベクトルからの漏れを抑える事ができる．

3.4 数値情報抽出

本節では，クエリに関連した数値情報を抽出について説明する．数値情報とは統計量，統計量の時間情報，統計量に対する比較表現の組み合わせである．地震の震度などの動向情報の場合，更に，一つ以上の地名も数値情報と共に追加される．

地名・震度情報の抽出

山田らが示しているように，新聞記事中の地震の表記には規則性があり,「◇震度６（烈震）　神戸」「奈良で震度４」「震度６を記録した神戸市」のように記されている [5].

そこで，構文解析器である CaboCha†を用いて規則に適合するものを選択し,「震度」と「地名」の組み合わせを抽出している．この抽出規則は「震度」だけではなく,「マグニチュード」にも適用している．

重要語の重み付け

クエリに含まれる重要語と，それら以外で関連文書集合に含まれる重要度の高い語も統計量の抽出に用いる．

重み付けは，関連文書中に含まれる各名詞，固有名詞，未知語に対して，酒井らが用いている重要語の重み付けをベースとした式 (1) を用いて行う [6].

$$w(t_i, S) = \frac{Tf(t_i, S)}{max_{i=1,\ldots,n} Tf(t_i, S)} \times \log \frac{|N|}{df(t_i, N)} \qquad (1)$$

t_i は各名詞，固有名詞，未知語，S は関連文書集合，$Tf(t_i, S)$ は文書集合 S における名詞 t_i の出現頻度，N は全文書集合，$df(t_i, N)$ は全文書集合 N において名詞 t_i を含む文書の頻度である．

この式を用いて重要度を割り当てることにより，抽出した関連文書中において各重要語に対して共通の重要度を用いることが可能である．

4 結果と考察

今回は，評価実験が不十分であるため，現時点での実験結果は記述せずに考察のみ述べる．

4.1 コーパスと実験対象

実験には，毎日新聞 95 年版と，読売新聞の 95 年版の 2 年分をコーパスとして用いた．コーパスには記事のキーワードなどの情報も含まれているが，本研究では記事索引番号と見出し，本文を利用した．

また，本研究では実験の主な対象として「兵庫県南部地震」に関連する記事を扱っている．

†http://chasen.org/~taku/software/cabocha/

4.2 パラレルコーパスの有効性

前述の「神津島」の例のように，パラレルコーパスを用いる事でデータスパースネスの問題解消が図れる．今回対象とした兵庫県南部地震の毎日新聞と読売新聞における関連記事の出現割合は

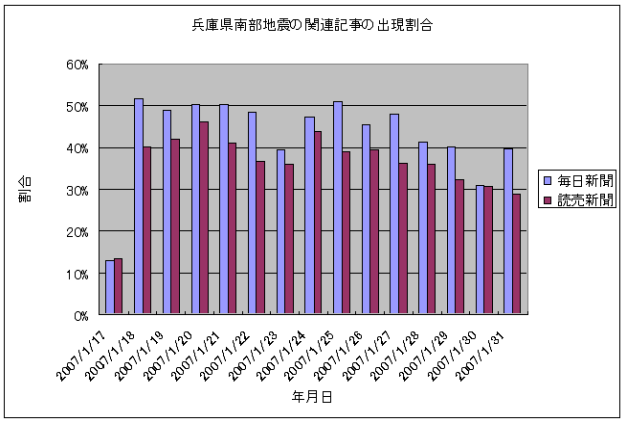

図 2: 兵庫県南部地震の関連記事の出現割合

図 2 のようになっており，全体的に毎日新聞の関連記事数が多いが，グラフの状態からも対応関係にある記事が十分に得られる事が分かる．しかし，発生直後という事もあり，この記事数は非常に多く，逆に不要な語を重要語として含んでしまうという問題も発生している．

4.3 見出しの規則性

兵庫県南部地震の場合，関連記事として抽出されたほとんどの記事の見出しが規則性のあるものであった．これもこの地震が大規模な災害であったために生じた特殊な状況であると考えられる．

毎日新聞では「阪神大震災　兵庫県南部地震－－余震３２回を記録」のように見出しの初めに「阪神大震災　兵庫県南部地震－－」が付けられており，読売新聞では「兵庫県南部地震　余震さらに１，２週間　Ｍ６級も」のように「兵庫県南部地震　」が付けられていた．

通常，新聞記事を対象とした文書抽出の重み付けでは，見出しや本文の 1 文目に特別に重みを与える事が多いが，このように見出しに規則性がある場合，重みを調整する必要がある．

具体的には，見出しの重みを重くしてあった場合,「阪神大震災　兵庫県南部地震－－東証株価にも地震の影響…生損保株が値下がり」という見出しを持つ，次の記事の類似度が高くなってしまう．

> 十七日午前の東京株式市場は，近畿圏の地震の影響が広がる中で，生損保株などを中心に値下がりし，東証の平均株価は先週末終値比百四十五円七十八銭安の一万九千百八十五円三十九銭で午前中の取引を終えた．
>
> 【毎日新聞 95.1.17】

この記事は，地震が要因となって統計量が変化したという意味では関連記事であるが，ここに書かれている統計量は全て地震とは直接関係の無いものである．そのため，地震を対象とした動向情報抽出では省かなければならない．

4.4 抽出可能な統計量

今回対象とした兵庫県南部地震は非常に大規模な災害であったために，抽出可能な統計量の種類は多岐に亘っており，震度や余震回数などの地震記事で得られる基本的なものから，仮設住宅の戸数や義援金の金額など非常に多くの統計量を得る事ができる．しかし，地震を対象として作成した統計量とその単位を登録した辞書には記述されていないものが多く，統計量として処理できないものも多かった．

4.5 文書要約

統計量以外の文書で提示すべき動向情報について，昨年度のシステムでは文書で表記された統計量の増減に関しては要約を行ったが，今回は関連記事の提示，パラレルコーパスでの対応記事の提示のみに留まっている．

地震を対象とした場合は文書で書かれている情報が非常に多く，本来文書要約を行うべきであるが，統計量の場合ほどの動向情報の抽出精度を得られないため，行えていない．パラレルコーパスについてはそれぞれの記事間で差が小さいので，まずこの記事間での要約が行えるようにする事が必要である．

5 終わりに

本稿では，パラレルコーパスを用いた新聞記事から空間的な動向情報の抽出と提示を行うシステムの提案を行った．実験と評価については本稿で全く扱っていないため，提案システムの客観的な評価を早急に行う必要がある．また，今回抽出した空間的な情報は文字と数値のグラフ化による動向情報の提示しか行っていないが，空間的な情報を地図にマッピングして提示する可視化は，非常に有用な情報提示の手段であり，実際に手法も提案されており，今後はこの可視化にも取り組みたい．

参考文献

[1] Tsuneaki Kato, Mitsunori Matsushita, and Noriko Kando: “MuST:A WorkShop on Multimodal Summarization for Trend Information”, in *Proceedings of the NTCIR-5 Workshop Meeting*, pp.556-563, 2005.

[2] 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，Vol.19，2005.

[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”, 情報処理学会自然言語処理研究会, 2004-NL-164 (15), pp.89-94, 2004.

[4] 曽我 真也，鈴木 宏哉，斎藤 博昭：“動向情報提示システムの構築”，動向情報の要約と可視化に関するワークショップ 第一回成果進捗報告会予稿集，pp.31-34，2006.

[5] 山田 隆志，中野 純，高間 康史：“タグ付きコーパスを用いた地震記事からの地理的動向情報の可視化”，言語処理学会第 12 回年次大会ワークショップ「言語処理と情報可視化の接点」論文集，pp.9-12，2006.

[6] 酒井 浩之，増山 繁：“ユーザとのインタラクションを導入した複数文書要約システム”，言語処理学会第 10 回年次大会発表論文集，pp.285-288，2004.