自然言語処理 125－6
（1998. 5. 29）

数値情報をキーとした新聞記事からの情報抽出

斉藤公一　迫田昭人　中江富人　岩井禎広　田村直良†　中川裕志
横浜国立大学 工学部　† 横浜国立大学 教育人間科学部

近年新聞記事など大量のテキストが電子化されているのにともない，テキストから重要な情報だけを抽出したいという要求が高まってきている．電子化されたテキストから重要な情報を抽出する技術は米国を中心に盛んに研究されており，高精度の抽出技術が確立されている．今までの情報抽出はテンプレートを埋めるために種々の複雑なパターンを作成し情報の抽出を行なっていたが，本報告で我々はドキュメント中に記述されている情報として，株価や人員の増減などの数値をともなう情報が重要であると考えた．そこで情報抽出のキーとして数詞に注目し，その周辺の言語パターンの解析により情報を抽出する．具体的には，これらの情報が文中でどのように現れるかを調査し，数詞の表す数値情報に関連する情報が記述されている言語表現パターンを発見し，それらのパターンをもとに，情報抽出を行うシステムを試作する．

Numeral Information Extraction from Newspaper's Articles

Koichi Saito, Akito Sakoda, Tomito Nakae, Yosihiro Iwai, Naoyoshi Tamura† and Hiroshi Nakagawa
Faculty of Engineering,Yokohama National University
†Faculty of Education and Human Science,Yokohama National University
{junkie@naklab, sakoda@tamlab, tommy@naklab, iwai@tamlab, tam@tamlab, nakagawa@naklab}.dnj.ynu.ac.jp

This paper describes an information extraction system which extracts numeric information that appears in newspaper's articles such as up/down of stock price, amount of employment, and so on. Numeric information is very important in articles to find movements of a specific domain like stock price of a certain company. To understand role of numeric word, we find lingistic patterns surrounding a numeric word. We develop a numeric information extraction system that is based on lingistic pattern matching.

1 はじめに

新聞記事などのテキストが電子的な媒体を介して流通・蓄積されることが一般的となって来ている．この膨大なテキスト情報を有効に活用することを目指してさまざまな研究が行なわれている．多量のテキスト情報から所望のテキストを捜し出すことが情報検索であり，所望の情報を有用な形に整形し提示することが情報抽出であるとされている．この情報抽出に関する研究は米国ではMUC(Message Understanding Conference)(MUC, 1996)において盛んに行なわれてきた(若尾, 1996a)．日本語を対象とした情報抽出に関する研究もMUCの多言語版であるMET(Multilingual Entity Task)においてさまざまな研究が行なわれており(福本, 下畑, 桝井, 佐々木 美, 杉尾, 1998)，名称認識などでは高い認識率を達成している．日本人名の認識としては(久光 丹羽, 1997)があげられる．

新聞記事の解析としては(西野, 落合, 木田, 乾, 桑田, 橋本, 1998)があげられる．ここでは，新聞記事の構造に注目して記事の一文目と以降にわけて解析を行なっている．対象は組織合併情報，新製品情報などであり，求める情報によって新聞記事の構造を解析し抽出パターンを作り直さなければならない．

また，(井出, 藤吉, 永井, 中村, 野村, 1997a)(井出, 永井, 中村, 野村, 1997b)においても新聞記事から字面のマッチングにより製品情報を抽出している．しかし，抽出する情報の種類毎に膨大な数のテンプレートを作成しなければなない．また，(松尾 木本, 1995)ではワイルドカードを用いた解析を行なっている．

その他の情報抽出として(長谷川 高木, 1998)があげられる．これは，字面マッチングにより日々交わされる電子メールからスケジュール情報を抽出するものである．

このような種々の研究成果がでてきているものの，文中に頻繁に現われ，かつ実用的にも重要な数値情報の抽出については大きな注意が払われてこなかったといえる．そこで，我々は数値情報に着目し，それに関連する情報を抽出しデータベース化することを試みた．以下，2節で研究の動機，システムの概念，3節でシステムの概要を述べ，4節でシステムの評価を行なう．最後に5節で今後の展望について述べる．

2 数値情報抽出

2.1 研究の動機

新聞記事などの大量のテキストデータから株価の動向や企業の雇用情報などを知りたい場合には，従来のキーワードマッチングに基づく情報検索，またそのための固有名詞の抽出などでは情報としては不十分な場合が多い．例えば，売り上げ高や株価，雇用状況などの現状や時系列的変化を知りたい場合には，「昨年度の売り上げは300億円に達した」「株価が100ドル上昇」「1000人規模での人員削減を行なう」などのような表現から有効な情報を抽出することが必要になる．このような情報を抽出する場合，数値情報に関連する情報の抽出を通じて数値情報を含んだ情報検索をすることが必須となってくる．

「ABC社」のような特定の会社，すなわち同一の実体，あるいは「ソフトウェア産業」のように同類の実体に対する記述が新聞記事には多く含まれる．それに付随する数値情報をデータベース化することによりその実体の変化(収益の増減など)を時系列として把握できる．この時系列変化の把握は，同一ないし同類の実体の動向調査などに欠くことのできない作業である．したがって，このような情報を記述するデータベースが情報抽出技術によって新聞記事などから自動的に形成できれば社会的に大きなインパクトを持つ．また一般的に新聞記事・技術論文等では文章中に数値が現われることが多く，文章中での数値とそれに関連する情報は重要度が高いと言える．

数値情報は形態素解析によって数詞を認識すれば，名詞・固有名詞に比べて認識

することが容易かつもれなく行うことが可能であり，確実に抽出できる．よって，上記の時系列変化を表すために必要な個別のひとまとまりの情報を取り出すための標識としてとらえやすい．

ただし，数値に注目しているだけでは，数値に関連しない重要な情報を落してしまうことになる．そこで，数値以外の情報は従来の情報抽出・情報検索技術により補うことも考慮する必要がある．

2.2 システムの概念

2.2.1 パターン駆動型情報抽出

従来の名詞・固有名詞抽出では単純なパターンに基づく抽出だけでなく複雑な構文解析などを行ない，目的情報の抽出を行なっているシステムも多く存在する．例えば(若尾, 1996b)では，英語新聞記事から固有名詞を抽出するために統語的な情報だけでなく，意味論，語用論のレベルまでの情報や固有名詞間での照応関係の情報などを活用している．しかし，複雑な構文解析は処理時間が長いこと，解析結果の曖昧さが大きいこと，などの問題から多くの情報を実時間で処理するには適さない．新聞記事など時々刻々増加する情報を迅速に処理するためにも解析に時間を必要とする構文解析の使用は避けるべきであると我々は考える．そこで本研究では我々はパターン駆動型でシステムを構築することを基本概念とする．

2.2.2 抽出する情報の形式

本研究は数値とそれに関係する情報を有意義な形式で抽出することを目的としている．数値情報の利用を考えると，記述される対象に対して数値がどのような情報を担っているのかが重要である．ここで対象とは，会社などの組織，特定の産業，売上高，株価，解雇人員などである．対象についての記述は二つの場合が考えられる．一つめは，対象の状態とその時点を表す時間情報である．二つめは，対象の行為とその時点を表す時間情報である．これらの表現の仕方が情報抽出におけるキーポイントになる．状態および行為はその対象の変化，状態の場合はその属性についての情報に関して，以下に記述する要素で抽出情報を表現できる．

1. 変化した対象(の名前)
2. 変化以前の状態
3. 変化後の状態
4. 変化量
5. 変化の定性的方向(増減など)
6. 変化の時刻に関する情報
7. 背景情報

ここで背景情報とは，対象が「株価」であれば会社名や産業名などである．つまり，対象に関連し，それを規定する情報である．以下では，これらの情報を抽出することを目的とした自然言語処理について述べる．

3 パターン駆動型情報抽出法

本システムはテキストを形態素解析システムJUMAN(松本, 黒橋, 宇津呂, 妙木, 長尾, 1996)を用いて形態素解析し，その結果にパターンを適用し，以下に述べるスロットの組(ユニット)を出力とするものである．

3.1 システム概要

図1に本システムの概要を示す．テキストを形態素解析し，その結果を利用して数詞情報，定性的表現の抽出を行ない，情報抽出パターンと照合しスロットを埋めることにより出力とする．

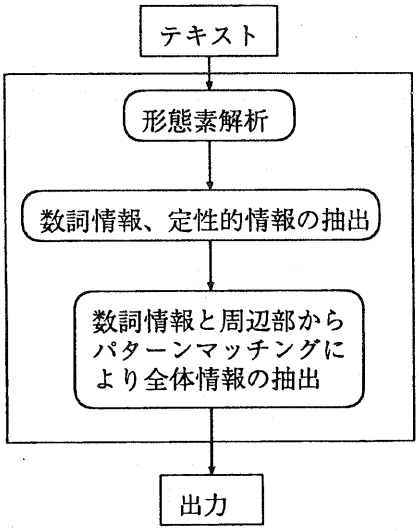

図 1: 情報抽出システム

3.2 出力形式

出力には，背景，name，結果，変化，もと，方向，時間, の七つのスロットからなるフレームを用いる．このフレームを，本研究では「基本意味ユニット」とよぶことにする．それぞれの内容を次の文を例として説明する．

例: 「事業の売却と買収を繰り返して，連結ベースの従業員数を三十二万人強から五万人削減し，九二年には二十七万人弱にした．」

name

変化した対象 (の名前) であり，数詞と係り受け関係にある名詞である．例えば「結果」などに入る数詞の単位が円であるとき，その単位でカウントされるような名詞が採用される．
例では「連結ベースの従業員数」

結果

記事内に現れる数詞，もしくは，ユニットの表す事象が動的である場合の変化した後の値である．name のスロットに入る名詞と係り受け関係にある．
例では「二十七万人弱」

変化

記事内の一文中に現れる数詞で，ユニットの表す事象が動的である場合の変化分を表す数詞である．記事内に現れてない場合に「結果」と「もと」から導き出すようなことはしない．あくまでも記事内に現れる単語だけで表現される．
例では「五万人」

もと

記事内の一文中に現れる数詞で，ユニットの表す事象が動的である場合の変化する前の値である．これについても，記事内に現れてない場合に「結果」と「変化」から導き出すようなことはしない．
例では「三十二万人強」

方向

記事内の一文中でに現れる名詞，動詞，形容詞などで，ユニットの表す事象が動的である場合の変化の方向性を示す．これについても，記事内の表現から方向性を推測するようなことはせずに，記事内に方向性を示すような単語が現れたときのみこのスロットが使われる．
例では「削減」

背景

記事内に現れる名詞の中で name に採用されなかったような名詞が入る．本研究においては name ほどは重要ではないが，見逃せないような情報のためのスロットである．
例では「事業の売却と買収」

時間

記事に描かれている出来事の起きた時期，時間，もしくはかかった期間などを表す．
例では「九二年」

3.3 抽出パターン

以下に実際に作成したシステムの抽出方法について述べる.

まず, 形態素解析結果から語を分類する. 分類は,

1. 数詞とそれに付随する単位の組
2. 時間を表す表現
3. 名詞の連結 (“の” での連結も含む)
4. 度合表現
5. 変化を表す表現
6. 助詞
7. その他

以上の7種類に分類する.

その後, 言語パターンに照らし合わせて, スロットを埋めていく. 以下に処理の過程を前述の例を用いて示す.

分類結果の例を図2に示す.

名詞{事業の売却と買収}:を:動詞{繰り返して}:,:名詞{連結ベースの従業員数}:を:数詞{値=三十二万,単位=人,度合=強}:から:数詞{値=五万,単位=人}:変化{削減}:動詞{し}:,:時間{値=九二,単位=年}:に:は:数詞{値=二十七万,単位=人,度合=弱}:に:動詞{した}:.

図 2: 品詞による分類の例

分類により一般化されたものを図3に正規表現の形で示すような抽出パターンに適用して抽出を行なう.

名詞:(に|で).*(を|が|は|も):
数詞:から:時間:変化

「name」に名詞, 「もと」に数詞,
「時間」「変化」が埋められる

図 3: 使用するパターンの正規表現の例

このようなパターンを適用しスロットを埋めると図4のような基本意味ユニットになる.

背景 事業の売却と買収

name 連結ベースの従業員数

結果 値 = 二十七万
単位 = 人
度合 = 弱

変化 値 = 五万
単位 = 人

もと 値 = 三十二万
単位 = 人
度合 = 強

方向 削減

時間 値 = 九二
単位 = 年

図 4: 抽出結果基本意味ユニットの例

4 システムの評価

本節では, 今回作成した数値情報抽出システムの評価実験を新聞記事を対象に人手で作成した正解集合をもとに行ない, その結果の検討をする.

4.1 試験システム

今回我々はシステムを2つのアプローチで作成した.

手法A できるだけ全てのスロットを埋めようとする

手法B 数詞毎にnameを見つけようとする

手法Aでは,定性的な情報を認識した後(図2の状態)でパターンを用いて可能な限りの名詞と数値の組み合わせを解候補として抽出し,数値の単位からnameを辞書を用いて推測する.推測によって絞り込まれた解候補の中から最も多くスロットを埋めるものを解として採用する.これに対して手法Bでは,パターンに優先順位が定義されており,その順位に従ってパターンが適用され数値に対するnameが抽出される.この時一度パターンが適用された数値を除外する.そして,スロットが確定していない数値が無くなるまでパターンの適用を続ける.なお図3のようなパターンを手法Aで119個,手法Bで120個使用している.

4.2 正解集合の作成手法

評価に用いる新聞記事は日本経済新聞１９９３年経済面の記事６００記事から抜粋した３００記事について人手で一文毎に正解ユニットを作成した.記事を一人５０記事ずつに分け作成した.

4.3 評価方法

抽出精度の評価方法として今回我々は,ユニットを単位とした正解率とスロットを単位とした正解率を求める.“name”については正解と部分マッチするものを考慮に入れる場合と正解と抽出結果が全く等しい場合とに分けて正解率を求める.ただし,“背景”については正解の厳密な定義が困難であることから,今回は抽出結果の評価は行なわない.ただし実際に検索には用いられるようにする.

4.4 評価結果とその検討

新聞記事での抽出精度を表1,表3(手法A)および表2,表4(手法B)に示す.

表 1: 大域的な抽出精度: 手法A

	全数	正解数	正解率(%)
完全一致	315	177	56.2
部分一致	315	200	63.5

表 2: 大域的な抽出精度: 手法B

	全数	正解数	正解率(%)
完全一致	235	122	51.9
部分一致	235	131	55.7

表 3: スロット毎の抽出精度: 手法A

スロット	全数	正解数	正解率(%)
name	387	254	65.7
結果	366	301	82.3
変化	38	30	79.0
もと	35	18	51.4
方向	63	52	82.5
時間	104	87	83.7
合計	993	742	74.7

表 4: スロット毎の抽出精度: 手法B

スロット	全数	正解数	正解率(%)
name	209	177	84.7
結果	194	165	85.1
変化	23	23	100.0
もと	15	12	80.0
方向	66	49	74.2
時間	92	74	80.4
合計	599	500	83.5

両手法においても，「name」の抽出精度が低いことがわかる．その理由としては形態素解析の失敗が大きな原因である．「結果」の誤抽出の原因は，商品名等に含まれる数字を数値と誤解してしまう場合が多いことである．「もと」の誤抽出の原因は，「価格はA円から」というものを抽出してしまっている．「方向」では，数詞から距離が遠い変化を表す語を採用してしまっている．「時間」では，「ゴルフ歴２０年」といった表現は本システムの「時間」スロットには適さない表現であるが採用してしまっている．

5 将来展望

ここでは，本システムを基にした将来のアプリケーションについて述べる．

5.1 数値に幅を持たせた検索エンジン

我々は数値に幅を持たせた検索エンジンへの本システムの組込みを考えている．数値に幅を持たせた検索エンジンとは，従来のターム検索とは異なり，

「従業員三千人のIT企業のここ
三カ月の株価動向を知りたい」

といった検索要求がなされた場合，従来のターム検索では各タームに関連する文章が検索されていた．しかし我々の考えるシステムでは実際に「三千人」「従業員」といった表記にマッチして検索結果とするのではなくて，自然言語でなされた質問を本システムで用いている情報抽出パターンを用いて解析し，文章から抽出されたデータベースとのマッチングをとり，その度合によって文章をランキングする．その際に，データベース上で質問文に現われる数値(例の場合「3000人」「3カ月」)に等しいものだけでなく，ある一定の規則に基づいて幅を持たせたマッチングを行なう．このことによって「二千九百人」などの表記でも，「三千人」との語のマッチングは無いが，「約三千人」として検索結果とすることが可能である．また「ここ三カ月」といった表記の意味を正確に把握することが本システムの手法を用いると可能である．

現在我々はこの検索システムを作成中であり，機会を改めて発表する予定である．

5.2 テキスト以外のデータベースとの連係

テキストに含まれている数値情報を抽出しデータベース化することにより，他のデータベースとテキストデータベースとの仲介役を本システムで作成したデータベースが数値情報を介して担い，テキストデータベースと他のデータベースとの親和性が高まる．

6 おわりに

テキスト中で重要な情報である数値情報に注目し，それに関連する情報を的確な形で抽出することを目的として数値情報抽出システムを作成した．時々刻々追加・更新される新聞記事を例にとって実時間での処理が可能であるパターン駆動型で情報抽出システムを作成し，その抽出精度の検証を行なった．本報告時点では抽出精度は決して満足できるものではないが，抽出パターンの検証・拡充を行なうことで抽出精度の向上が期待できる．本システムを用いた将来の応用システムについても述べた．また新聞記事以外に「農業普及データ」への適用も現在進行中である．

謝辞

本研究の一部は農林水産省の一般別枠研究「増殖情報ベースによる生産支援システム開発のための基盤研究」の一環として行なわれた．

参考文献

井出裕二, 藤吉誠, 永井秀利, 中村貞吾, 野村浩郷 (1997a). “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.” 情報処理学会研究報告 97-NL-118-2, 情報処理学会.

井出裕二, 永井秀利, 中村貞吾, 野村浩郷 (1997b). “単一項目テンプレートによる新聞記事からの製品情報抽出.” 情報処理学会研究報告 97-NL-122-10, 情報処理学会.

西野文人, 落合亮, 木田敦子, 乾裕子, 桑田和佳子, 橋本三奈子 (1998). “トップダウンな解析に基づく情報抽出.” 情報処理学会研究報告 98-NL-124-13, 情報処理学会.

長谷川隆明 高木伸一郎 (1998). “電子メールコミュニケーションにおけるスケジュール情報抽出.” 情報処理学会研究報告 97-NL-123-10, 情報処理学会.

久光徹 丹羽芳樹 (1997). “辞書と共起情報を用いた新聞記事からの人名獲得.” 情報処理学会研究報告 97-NL-118-1, 情報処理学会.

福本淳一, 下畑光夫, 桝井文人, 佐々木 美樹, 杉尾俊之 (1998). “パターン処理に基づく情報抽出システムの概要.” 言語処理学会第四回年次大会発表論文集 A4-4, 言語処理学会.

松尾比呂志 木本晴夫 (1995). “抽出パターンの階層的照合に基づく日本語テキストからの内容情報抽出法.” 情報処理学会論文誌, **36** (8), 1838–1844.

若尾孝博 (1996a). “英語テキストからの情報抽出 ＭＵＣ第６回大会の参加報告.” 情報処理学会研究報告 96-NL-114-12, 自然言語処理研究会, 情報処理学会.

若尾孝博 (1996b). “英語新聞記事からの固有名詞自動抽出技術.” 情報処理学会研究報告 96-NL-115-9, 自然言語処理研究会, 情報処理学会.

松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真 (1996). 日本語形態素解析システム JUMAN. 京都大学工学部 長尾研究室, 奈良先端科学技術大学院大学 松本研究室.

MUC-6 (1996). *Proceedings of the sixth Message Understanding Conference, Columbia,Maryland,U.S.A. 1995.*