新聞記事内の企業の抽出と識別

伊吹 潤, 西野 文人
ibuki.jun,Nishino@jp.fujitsu.com
富士通研究所ドキュメント処理研究部

1 はじめに

企業データベースでは信頼度の高い情報が得られる一方で、限定された項目の情報しかない、最新の情報がないといった問題がある。一方で新聞記事は様々な情報を日々供給しており、両者をうまく組み合わせることでユーザーに、より完全な情報を提供することができるだろう。

我々はある企業について検索した際に企業データーベース内のデータと共に新聞記事データベース内のその企業の動向についての情報を提示するシステムを開発している。ここでは新聞記事中に記載されている企業の情報を正しく認識すること、新聞記事内の企業記述と企業データベースのエントリとの対応関係を正しく識別すること、更にその中から指定された企業の活動について記述されている記事だけを抜き出すことが求められている。

本稿ではこれらの課題に対する我々の考え方、及び処理手段について述べ、システムの評価実験の結果を示す。

2 提示された課題についての検討

2.1 新聞記事中の企業の認識

新聞記事中の企業を正しく認識するためにはまず企業名として使われる名詞を普通名詞として使われる名詞を区別して取り出すこと（ex. 企業名としての「アクセス」と「アクセス・タイム」と区別する）が求められる。

又対象企業名の範囲を正しく切り出すことも当然ながら重要である。これは一般に考えるよりも大変である。例えば形態素解析ベースの処理では富士通電装という会社を富士通として切り出したりすることがよくある。

2.2 企業データベースとの照合

曖昧さの扱い 新聞記事中の企業情報と企業データベース中のエントリ間の対応をつけるためには同名異企業間の区別をすることが必要となる。例えばアクセスという名前に対応する企業は帝国データバンク中で100件以上存在する。

一般に無名の企業の場合は代表者や所在地等の情報が同時に書かれているのでこうした情報を利用して識別をすればよいのだが、一方で有名企業の場合は名称のみの記述の場合が多く、却って大きな問題となる。例えば日本電気は日電として知られる有名企業とは別の会社が存在しており、日本電気とだけ記述された時にどれを選ぶかが問題となる。

別表記への対応 現在、ＣＩ等で広報で使用する名称が登記された正式名称と異なる企業は少なくない。例えば小松製作所は製造業のみにイメージを固定されたくないということでコマツというカタカナ名を使用している。他には店舗チェーンの方が企業名よりも有名となってしまったユニクロ（企業名はファースト・リテイリング）の例もある。

又新聞記事中では文字数の節約のため、しばしば正式商号以外の表記（例えば日本アイ・ビー・エムは新聞中ではしばしば日本ＩＢＭと略記される）が用いられる。

こうした大きな違い以外にも表記の揺れ（新旧の字体の違い、カタカナ表記の揺れ）は企業名にもしばしば現われる。これらは検索における再現率を下げる大きな要因となり得る.

関連記事の選択 企業名は新聞記事中の様々なところに出現する。例えばスポーツ選手や訃報記事中に人物の所属先として出てきたり 、中心的なトピック

の後に参考情報として出てきたりする。しかし利用者がある企業について調べるために記事を見る場合に求めるのは、対象企業自身が中心となった記事であり、上述の記事はほとんどの場合不必要となる。

記事の重要部分を抽出するためには表層的な構文[1]や出現頻度等を利用した例があるが、十分な精度が出ているとは言い難い。我々は、記事の中心的なトピックにおいて対象企業が主体となった活動、企業トップの交代、企業の属性（業績、製品など）に関する情報が記載された記事を選択することにした。例えば図1の記事では「ニチワ」と「三菱商事」の2つの企業が記載されているが、「ニチワ」の製品に対する取り組みが記事のトピックとなっているのに比べ、「三菱商事」は周辺的な情報である。従ってこの記事は「ニチワ」の関連記事としては適当だが、「三菱商事」の関連記事とはならない。

3 企業情報の抽出・識別システム

ここではシステムの構成を図2に示し、各部の処理について述べる。

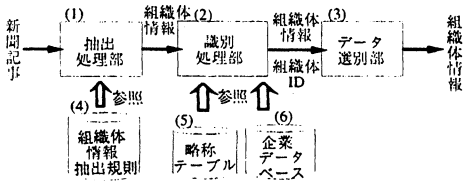

図 2: 企業情報の抽出・識別用システムの構成

3.1 抽出処理部

抽出処理部では新聞記事の構成、出来事の記述方法、企業名の構成等に関する知識を利用して企業名と他の項目情報（所在地、代表者等）を抽出する。企業情報と共に、どのような手がかりを利用したかについての情報も保存しておき、最後に関連記事の選択の際に利用する。

ここでは抽出のためのヒューリスティックスを組織体情報抽出規則として処理エンジンとは別の部分にまとめることで保守性や了解度の向上を図っている。ヒューリスティックスは[2]で示されたものを更に発展させたものであり、以下のような内容をもつ。

- 企業の経済行動等の出来事を認識し、それらの主体となる部分から企業情報を抽出する。
- 新聞がもつ新出の会社を記述するパターン（代表者や所在地等を説明する書式）を手がかりとして企業情報を抽出する。
- 名前自体の持つ特徴等（ex. 「～社」、「～協会」等）によって企業名を抽出する。

上記のヒューリスティックスによって抽出された企業名だけを企業データベースとの照合の対象とすることによって普通名詞が企業情報中に紛れ込むことはなくなる。又企業名のリストとのマッチングによって切りだしを行なう訳ではないので、富士通電装から富士通を切り出すようなことはない。

3.2 識別処理部での処理

識別処理部では抽出された企業の名称や所在地、代表者、電話番号等の情報を利用して企業データベース中のエントリとの照合を行ない、企業を特定する。これによって同名の企業が多数存在するような企業でも代表者名や電話番号等の情報を利用すれば、一意に企業を特定できる。

ここでは企業データベース、略称テーブルの二種類のデータベースを持っている。略称テーブルは名称と企業データベースのエントリの対応表であり、名称をキーとした検索を行なうことができる。識別処理ではまず略称テーブルの検索によって得た候補とのマッチングを行ない、それに失敗した場合に企業データベース本体の検索を行なうようにしている。

略称テーブルの検索は2つの役割を持っている。一つは正式名称以外の略称、別称によって記述された企業情報の識別である。このため我々は一般的に流通している略称データを約1万件、更に新聞での略称を企業データベース中の正式名から機械的な操作によって生成し、登録している。

もう一つの役割は有名企業の企業記述の識別時の曖昧さを解決することである。日本電気のように同名の別企業が存在する場合に備えて略称テーブルには東証大証の一部二部に上場した企業について、正式名と対応する企業エントリを登録し、名称のみの情報に対しては略称テーブル中のエントリのみが得られるようにしている。

【静岡】包装材料のニチワ（静岡県沼津市岡宮１０８１の６、阿部留松社長、０５５９・２３・４５６７）は、荷崩れ防止エアバッグ「エアダンネージ袋」で攻勢をかける。... 販売面では、... ３年前に結んだ三菱商事との総代理店契約を今月中に解消し、専門の包装資材商社に販売を任せる。...

図 1: 日刊工 2001.9.12 の記事（一部略）

企業情報と企業データベース中のエントリとの照合に当たっては表記の揺れ等に関する知識を利用した柔軟なマッチングを行なうことによって企業データベース中の表記と新聞記事中の表記の微細な違いに対処している。又所在地の表記方法の構造的な違い（ex. 番地間のセパレータの違い、「字」の扱い）等については基本的に企業データベースの書式に向けた標準化をした後でマッチングを行なうことで対処をしている。

更にこの時点で対応するエントリに曖昧さがある場合は、各候補エントリに対して会社の資本金の割合に対応したスコアをつけている。一般には資本金額の大きさが企業の知名度に対応するので、スコアが大きい企業エントリが対応する可能性がより大きくなるはずである。

3.3 データ選別部での処理

ここでは主に企業情報の抽出時に抽出された各情報によって企業情報の重要度の評価値を次のような基準によって計算する。

- 記事のトピック文において企業の経済活動の主体となった企業、トップの動向についての記事における企業、業績についての報告記事における企業を最重要と評価する。
- 上記以外の場合は個々の手がかりによる点数付けを行ない、各評価値を集計することで全体としての重要度を評価する。（例えば文の主語であるか、後置情報をもつか等の手がかりについて個々の評価値を持つ）

  これはトピックの認定に失敗した場合にも、状況証拠の積み重ねによって関連記事かを判断するためである。

4 システムの性能評価

4.1 キーワード検索結果との比較

日刊工業新聞の３カ月分のデータベースに対して企業名をキーワードとして指定した検索を行なった検索結果を基に、その中で実際に企業が出現した記事（出現記事）、更にその中での関連記事について人手で調べてみた結果をまず示す。関連記事を選択した基準は第２節で述べた通りである。

対象企業	富士通	三菱商事	アルファ	アクセス
KW検索[1]	375	217	51	486
出現記事[2]	301	211	2	10
関連記事[3]	48	23	2	6

1 キーワード検索によって得られた記事の総数

2 1の中の対象企業が出現した記事の件数

3 2の中の対象企業の関連記事の件数

表 1: キーワード検索結果中の関連記事の調査

2の結果を見るとキーワード検索の結果に望まない記事（検索ゴミ）を含むかは企業によって大きく違うことが判る。特に「アルファ」、「アクセス」は一般名詞との混同が多く現われる。一方富士通や三菱商事の場合の検索ゴミの割合ははるかに少ない。全く0にならないのは社名を名前の一部に持つ関連会社との混同のためである。

3の関連記事の件数をみると300件以上の出現記事のあった富士通も富士通自身の動向についての記事（関連記事）はその2割にも満たないことがわかる。同様に三菱商事の場合も製品の販売元としての記述が多く三菱商事自身の動向についての記事の割合は一割強といったところなのがわかる。

4.2 システムの処理結果との比較

先ほどの4つの企業について企業情報の抽出・識別を行なって実際に得た結果を人手によって得た関連記事と比較した結果を示す。

対象企業	富士通	三菱商事	アルファ	アクセス
関連記事[1]	48	23	2	6
出力結果[2]	40	19	2	6
適合率	40/40	19/19	2/2	6/6
再現率	40/48	19/23	2/2	6/6

表 2: 関連記事の処理性能の評価

1 対象企業の関連記事の件数

2 システムが関連記事と認定したものの件数

これから関連記事の抽出処理の適合率が非常に高く、システムによって選ばれた関連記事中に誤った記事が入り込むことは珍しいということが判る。

抽出パターンへの抜け等の原因によって関連記事抽出の再現率は7、8割の水準であるが、関連記事として全ての記事を提示することはどのみちできないので、我々の使用目的には十分な性能と考えている。

4.3 関連記事の抽出機能の評価2

先の評価ではキーワード検索の結果を母集団としたために関連記事の抜け（再現率の低下）の点での不安があった。このために日刊工一日分の記事の全体についての関連記事のチェックをしてみた。

新聞名	適合率 (%)	再現率 (%)
日刊工	98(115/117)	80(115/142)

表 3: 日刊工 2001 年 4 月 2 日分の記事（231 記事）に対する評価

ここでも関連記事抽出の適合率、再現率については、企業を固定して行なった評価とほぼ同じであり、再現率の点で大きな低下はないと考えられる。

4.4 抽出処理機能の性能評価

ここでは日刊工3カ月分の企業情報を対象にして識別機能のみの評価を行なってみた結果を以下の表に示す。ここでは官公庁や外国企業等、対象の企業データベース（帝国データバンク）の対象とならないものは最初から除いてある。

項目	一意決定[1]	未発見[2]	曖昧[3]	その他
割合	85%	11%	3%	1%

表 4: 識別機能の評価結果

1 対応するエントリを一意に決定

2 対応するエントリが見つからず

3 対応するエントリが複数ある

未発見のエントリの割合が一割以上を占めている。これらは人手でのチェックでもデータベース中に対応エントリは見つからなかった。我々は、これらを会社としての登記を行なっていない家族経営のものと考えており、データベースのエントリに対する関連記事の提示という機能からは問題はないと判断している。

対応するエントリが複数あるもののほとんどは名称のみの記述であり、スコアによる選択によってこれらの半分程度を正しく識別できるという見込みを得た。残っている部分は抽出の際の切りだし誤りや未対処の種類の表記揺れによるものである。

5 おわりに

現在, 企業が提供する Web サイトの充実は著しく, プレスリリースや製品のサポート情報の点では新聞より充実していると言えよう。ところが技術情報を企業の枠を越えて検索しようとすると、企業毎に語彙や情報の整理の仕方が違い, 非常に検索ゴミが多くなるのが実情である。

今後はトピックを技術, 製品などのテーマに広げると共に,Web 上の情報を含めて種々の情報源からの情報を総合してユーザーに利用しやすい形態で提供することを目標として検討を行なっている。

参考文献

[1] 横山晶一他：“主題・焦点のスコアを用いたキーワードの抽出”, 言語処理学会第7回年次大会 (2001)

[2] 落谷 亮：“組織名抽出のための知識収集”, 言語処理学会第5回年次大会 (1999)