組織名抽出のための知識収集

落谷 亮
富士通研究所
ochi@flab.fujitsu.co.jp

1 はじめに

新聞記事ＤＢサービスなどで利用可能なテキストデータの量が増えるにつれて、テキスト中の組織に関する情報を識別し効率良く扱うことの必要性が以前にも増して高まっている。

例えば、ジー・サーチ [G-search] の提供するデータベースでは、一般紙、専門紙合わせて 16 種類の新聞記事の横断検索が可能であり、さらには、 100 万社レベルの収録数を持つ企業情報データベースの検索サービスも提供している。このようなデータベースサービスにおいて、新聞記事に代表されるテキストデータと企業情報データの情報を密に連携させることは、データベースの利用可能性を大きく広げることになる。

新聞記事テキスト中の企業についての記述と企業情報ＤＢを連携させるには組織名抽出が必要になるが、対象とする組織数が大規模になるにつれて、組織名を固有名詞辞書に蓄えれば済むというような簡単な問題ではなくなる。

本報告では、これらの大規模な組織情報の収集に対するアプローチとして、新聞記事における組織名固有の表現パターンを用いて組織名辞書を構築し、その組織名辞書を用いた組織名抽出を行なうとともに、パターン照合と組織名辞書による抽出の両面での処理を試みた結果について報告する。

2 組織情報抽出

組織名の具体的な抽出実験について述べる前に、組織名抽出または組織情報抽出のにより解決すべき課題として我々が認識している問題について簡単に述べ、組織名抽出での必要条件について整理してみる。

1. 新聞記事検索における絞り込み

   ある企業に関する記事を読みたい際に、現在のデータベース検索サービスでは、企業名で検索することが出来るが、企業名が一般の名詞句表現と重なる場合（「ネットワーク」、「デジタル」等）や、同名の企業が存在する場合（「株式会社大和」、「株式会社サンコー」等は同名が数百存在する）には、検索結果には本来必要だったもの以外のデータが混在する。

   - 組織名と一般名詞句の区別が必要。
   - 同名組織の識別が必要。

2. データベースのリンク（記事対記事、記事対企業情報）

   ある記事を参照中に見つけたのと同じ組織についての関連記事を読みたい場合、現状では記事検索しか方法は無く、 1の例と同じく、必要な情報だけを簡単に手に入れることは出来ない。

   企業の経営状況や株価などは企業情報に現れるが、新製品の発売や技術的な発表などは新聞記事に現れる。他社動向や業界動向などの調査で両方のデータを扱う場合を考えると、テキスト中の企業が識別され、企業情報と同時に扱えることで、マルチソースＤＢの有効性が容易に活かせるようになると考えられる。

   また、記事情報と合わせて企業情報も入手したい場合、企業情報データベースも利用可能であるが、代表的な企業情報プロバイダである帝国データバンク [帝国] では 110 万社、東京商工リサーチ [東商工] で 100 万社というように情報数が大規模であり、その結果、同名の企業（特に人名から由来するような企業名など）の情報が他数収録されており、自分の知らない会社を

探したい場合などは、検索によるのは非常に困難である。

- 記事本文の組織名と企業情報の対応付けが必要。
- 100 万社規模の組織名抽出では高再現率が要求される。

3. 過去の情報の利用

企業情報データベースからは最新の企業情報のみが入手可能であるが、新聞記事には過去の企業活動や事件の情報も、その当時の最新情報として書かれている。過去の企業情報を調べるには、当時の記事に書かれた名称や代表者名などの知識が必要になる。

- 代表者や企業名は時間と共に変わる情報であり、過去の情報の蓄積が必要。記事固有の表現（略称など）も蓄積が必要。

以上、組織情報抽出の応用イメージと必要とされる処理の要件を挙げた。まとめると以下の点が抽出処理には重要なポイントとなる。

1. 組織名と一般の名詞句の識別。
2. 同名の組織の識別。
3. 高再現率の実現。
4. 時間に伴う情報の変化への対処。
5. 略称などの新聞中での表現の処理。

これらの要件を満足させる情報抽出システムの実現に向けて、ここでは、以下の観点からの実験を行なった。

1. 組織名辞書の自動構築。
2. 過去の記事からの組織情報抽出。
3. 一般の名詞句と処理上紛れ易い組織名の識別。

3 抽出実験の概要

今回は、以下の一連の抽出実験を行なった。

- 既存の企業名データによる抽出。
- パターン照合に基づく抽出による組織名、その他の組織情報の抽出。
- 自動抽出結果を利用した組織名抽出。
- 処理誤りの多い組織名の判定処理。

実験 1 では、「CDROM 日経会社情報'98」で提供されている会社名データ 3356 社を組織名辞書として利用し、単純な文字列照合による抽出精度を求めた。評価の対象文書としては、日刊工業新聞 [日刊工] の記事 1000 記事を選び、その一文目に対し組織名部分を正解作成者にタグ付けして貰った結果を利用した。なお、抽出精度の測定には IREX[関根] の採点プログラムを利用した。

実験 2 では、パターン照合 [西野] による簡単な抽出処理プログラムを作成し、実験 1 の正解セットに対して抽出精度を測定した。

さらに、正解セットとは別に約 40 万記事の日刊工業新聞を対象に、組織名、代表社名、所在地、電話番号、その他の情報を抽出した。日刊工業新聞は専門紙であるため定型情報表現が多くみられ、パターン照合よる抽出には向いた題材と考えられる。

実験 3 では、パターン抽出で得られた組織情報のうちの組織名を用いて組織名辞書を作成し、作成した辞書を利用し実験 1 と同じ正解コーパスに対し辞書を用いた抽出実験を行なった。辞書の規模を変化させ辞書の規模と抽出精度の関係を測定した。

さらに、パターン抽出と辞書の併用による抽出を行ない辞書の規模と抽出精度の関係を測定した。

実験 1 及び実験 3 の抽出結果をみると、実験 1 では「アップ」のように一般の名詞と重なる企業名が誤抽出の原因になっており、実験 3 では、それに加えて自動収集による組織名でないデータが誤抽出の原因となっている。これらの誤抽出は特定のエントリに非常に多く出現するという性質を持つ。

そこで、実験 4 では、パターン抽出による適合率の高い抽出結果と文字照合による抽出結果の間で、抽出された組織名の頻度分布の差を利用して、精度低下の原因となっている組織名を選別し辞書から除去し、再度抽出実験を行なった。

4 実験

4.1 実験 1

図 1に原コーパスとテキストと正解のタグ付きコーパスを示す。今回は一文目だけの正解データを用いているが、記事全文の正解データを用意するのが本来は望ましい。

抽出採点プログラムによる採点結果は図 2の通りである。

```
<DOC>
<DOCNO>1</DOCNO>                                    # 原文
<TEXT>
  日本光電は在宅医療（テレ・ケア）事業に本格参入する。
</TEXT>
</DOC>

<DOC>
<DOCNO>1</DOCNO>                                    # 正解
<TEXT>
  <組織名> 日本光電 </組織名> は在宅医療（テレ・ケア）事業
に本格参入する。
</TEXT>
</DOC>
```

図 1: 原文とタグつき正解

再現率	適合率	F 値
46.13	61.87	52.85

図 2: 既存データによる辞書照合の抽出スコア

4.2 実験 2

照合に用いたパターンは、「～組織名（社長～、…）」の形を骨格にした派生系のパターンである。

このパターンを用いて、実験 1 と同じ評価コーパスに対し抽出を行なった結果は図 3の通りである。

再現率	適合率	F 値
36.18	99.04	53.00

図 3: パターン照合の抽出スコア

さらに、同じ抽出パターンを用いて日刊工業新聞の 40 万記事全文に対して処理を行ない 25 万件の結果を得た。図 4に抽出された組織情報を示す。図では、時間に対し連続して現れる情報をまとめてあり、2 カラム目は情報の掲載された時期を示している。この情報は、例えば社長の交代のような、時間の経過に伴う組織情報の変化を示している。

4.3 実験 3

実験 2 で得られた情報は、組織名相当部分として約 5 万 5 千件の情報を持つ。この抽出結果を実験 1 の企業名データに追加して組織名辞書として利用する。

この場合、組織名辞書の組織名件数や見出し語の性質に抽出結果は左右されることが推測されるので、組織名件数を 3350、 5 千、 1 万、 3 万、 5 万、

富士通	
	掲載時期 (頻度)
代表者 (山本卓真: 社長)	1987/06/01-1990/06/27(517)
代表者 (山本卓眞: 社長)	1987/08/19-1990/02/16(13)
代表者 (関澤義: 社長)	1990/06/29-1996/10/23(75)
代表者 (関沢義: 社長)	1990/07/02-1997/03/27(1315)
電話 (045-475-1956)	1996/05/11-1996/05/11(1)
電話 (03-3216-3211)	1996/08/03-1997/01/18(2)
電話 (0120-894321)	1996/04/20-1996/04/20(1)
電話 (03-3216-8035)	1996/01/29-1996/01/29(1)
市内石州府地内に工場進出が決まっている	1988/09/28-1988/09/28(1)
わが国を代表するコンピューターメーカー	1988/10/13-1988/10/13(1)
大阪大学の浜川圭弘教授らの研究を基	1989/03/02-1989/03/02(1)
キヤノン	1991/03/15-1991/03/19(2)
営業支援などの事務や工場の生産現場に従事している短大・高卒の女性社員を対象	1993/03/11-1993/03/11(1)
（以下略）	

図 4: パターン抽出による企業情報

全件 (5 万 5 千) と変えて抽出は行ない、先と同じ正解セットに対する抽出精度を計算する。なお、 3350 件は実験 1 との比較のために実験に含めた。

辞書照合による抽出を行なった結果を図 5に示す。

辞書数	再現率	適合率	F 値
3350	60.74	70.55	65.28
5000	63.38	47.89	54.57
10000	68.49	40.56	50.95
20000	72.10	32.21	44.52
50000	70.07	18.97	29.85
全件	69.72	18.33	29.03

図 5: 獲得組織名による抽出（辞書）

図 6に示すのは、組織名辞書照合とパターンによる処理を組み合わせて処理した場合の抽出精度である。組み合わせ方としては、パターン処理で組織名の抽出できなかった文に対し辞書照合を行なうという順で処理している。

辞書数	再現率	適合率	F 値
3350	81.51	85.66	83.54
5000	82.13	82.57	82.35
10000	82.13	70.63	75.95
30000	80.63	51.66	62.98
50000	79.05	37.94	51.27
全件	78.35	36.64	49.93

図 6: 自動獲得辞書による抽出（パターン＋辞書）

4.4 実験4

実験 4 では、先の 40 万記事のコーパスに対し辞書による抽出とパターンによる抽出を行ない、その結果から組織名の出現度数の順位を計算し比較した。

それぞれの抽出結果における組織名頻度は、元々のコーパスでの出現頻度順位と相関があるとの仮定に基づき、辞書抽出結果のうち出現頻度順位が異常に高い組織名は抽出誤りの元と判定出来るという考えた。

二つの抽出方法で結果の組織名総数は異なるので、組織名毎の出現頻度順位を組織名の異なり総数で割り、順位を 0 ～ 1 間の値に正規化し、(辞書抽出による値－パターン抽出の値) を評価値として計算した。値が小さいものから順に並べた組織名一覧を図7に示す。

この上位の 88 件の組織名を先の辞書から削除した後に抽出を行ない、その結果を採点したものを図8に示す。

組織名	評価値
アップ	-1.00
高速	-0.99
スペース	-0.99
ジャパン	-0.98
長大	-0.97
テック	-0.97
キング	-0.95
ジャパン	-0.95
中部	-0.95
大和	-0.94
KDD	-0.94
(以下略)	

図 7: 順位差のによる組織名一覧

辞書数	再現率	適合率	F 値
3350	80.55	91.78	85.79
5000	81.16	89.51	85.13
10000	82.66	81.65	82.15
30000	83.36	69.28	75.67
50000	82.92	58.40	68.53
全件	83.01	57.36	67.84

図 8: 削除後の抽出結果 (パターン＋辞書)

実験では、パターン抽出の結果を仮想的な正解と考え、辞書抽出の結果の比較対象としたが、現在のパターン抽出で扱えない組織名、例えば、通産相、大蔵省などの省庁、KDDなどは出現パターンが異なり、うまく扱えていない。

5 まとめ

組織名抽出処理として、辞書照合による抽出とパターン照合による抽出の組合せ処理の結果が良いことを示した。また、一般の名詞句表現と頻繁に照合し誤りの元となっている組織名を検出し辞書から除くことで、適合率を向上出来ることを示した。

ここで用いた評価用正解は 1000 文という小規模なものであり、抽出結果を利用して、さらに大きな正解データを作成していく必要があると考えている。

今回は扱わなかったが、組織名の略称について考えると、略称の付け方自体が規則的なもの (株式での略称) から愛称まで様々な種類があり、略称の参照先の識別なども含めて処理を考えるとさらに問題は複雑になる。略称がどのように使われているのかを、正解セットの拡張と合わせて多量のテキストデータから分析するなども今後の課題である。

今回の実験で抽出した企業情報は、それ自体が参照用の情報としても利用可能であるし、また、実際の利用を考えると重要な組織識別のための情報としての活用が今後の課題である。

最後に、貴重な新聞記事データを利用させてだいた日刊工業新聞社に感謝の意を表したい。

参考文献

[西野] 西野 文人, 橋本 三奈子, 落谷 亮
テキストからの用語とその定義文の抽出 言語処理学会第5回年次大会, A2-5, 1998.

[G-search] G-search 企業向けマルチメディアデータベースサービス
株式会社ジー・サーチ (http://www.g-search.or.jp)

[日刊工] 日刊工業新聞
株式会社日刊工業新聞社 (http://www.nikkan.co.jp)

[帝国] 株式会社帝国データバンク (http://www.tdb.co.jp)

[東商工] 株式会社東京商工リサーチ (http://www.tsr-net.co.jp)

[関根] 関根 聡, 井佐原 均
IREX: 情報検索、情報抽出コンテスト, 情処研報, 98-NL-127-15, pp. 109-116. 1998.

[日経] CDROM 日経会社情報'98 秋号
日本経済新聞社, 1998.