トランスデューサによる日本語固有表現抽出

佐々木 裕
NTTコミュニケーション科学基礎研究所

1 はじめに

情報抽出に関する研究は、Message Understanding Conference を中心にしてこの10年間に着実な進歩を遂げてきた。情報抽出の精度の向上のためには、基本ツールとして利用される固有表現抽出ツールが、高い精度を持つことが重要である。最近、IREX コンテスト [4] の企画により、日本語テキストについても、固有表現抽出ツールを同じ条件のもとで定量的に比較できる環境が整ってきた。このような環境の中で、日本語の固有表現の抽出の研究が盛り上がる機運がうかがえる。

本稿では、まず、情報抽出システムでの利用を考慮した固有表現のタグの定義の拡張[1]について述べる。拡張されたタグに従って、現在3000記事の正解タグを作成中であり、その過程で明らかになった判定基準について実例をまじえながら解説する。次に、拡張されたタグを付与する固有表現抽出ツールを、単純な置換ルールである有限状態トランスデューサ（以下、単にトランスデューサ）の集合により実現する方法について述べる。

2 IREX の固有表現タグ

まずは、ベースとなっている IREX の固有表現タグを簡単に紹介する。詳しくは [1] や IREX のホームページ [4] を参照していただきたい。IREX の固有表現タグは、大きく「固有名詞的表現」「時間表現」「数値表現」からなり、それぞれの細分類は以下のようになっている。

- 固有名詞的表現
  - ⟨ORGANIZATION⟩ 組織名、政府組織名
  - ⟨PERSON⟩ 人名
  - ⟨LOCATION⟩ 地名
  - ⟨ARTIFACT⟩ 固有物名
- 時間表現
  - ⟨DATE⟩ 日付表現
  - ⟨TIME⟩ 時間表現
- 数値表現
  - ⟨MONEY⟩ 金額表現
  - ⟨PERCENT⟩ 割合表現

3 タグの拡張

固有表現抽出結果の情報抽出システムでの利用を考慮して、以下のような固有表現タグの拡張を独自に行なった。特に、現在新聞等の製品情報記事から主要情報を抽出するシステム [2] の開発を行なっているため、単位を伴う数値表現の種類を重点的に追加した。また、「同社」といった照応表現を扱うため、照応関係のタグも追加した。人名の抽出の精度向上、および情報抽出における人名のクラスの判定のために、固有表現ではないが、補助的なタグとして役職・敬称タグを追加した。今後、さらにタグの追加、修正を行なう予定であるが現時点での拡張分のタグの一覧は以下の通りである。

- 時間表現
  - ⟨PERIOD⟩ 時間間隔表現
- 固有名詞的表現
  - ⟨PHONE⟩ 電話番号表現
  - ⟨EMAIL⟩ 電子メールアドレス表現
  - ⟨URL⟩ URL表現
- 数値表現
  - ⟨LENGTH⟩ 距離表現
  - ⟨SQUARE⟩ 面積表現
  - ⟨VOLUME⟩ 体積表現
  - ⟨NORGANIZATIONS⟩ 組織数表現
  - ⟨NPEOPLE⟩ 人数表現
  - ⟨NLOCATIONS⟩ 地名数表現
  - ⟨NARTIFACTS⟩ 固有物数表現
  - ⟨AGE⟩ 年齢表現

[1] この拡張は我々独自に行なった拡張であり、IREX の固有表現抽出コンテストで用いられるタグの拡張を意味しているわけではない。

- ⟨POINT⟩ 点数表現
- ⟨SPEED⟩ 速度表現
- ⟨CSPEED⟩ 通信速度表現
- ⟨FREQUENCY⟩ 周波数表現
- ⟨RESOLUTION⟩⟩ 解像度表現
- ⟨VOLT⟩ 電圧表現
- ⟨CURRENT⟩ 電流表現
- ⟨WATT⟩ 電力表現
- ⟨WEIGHT⟩ 重量表現
- ⟨PULSE⟩ パルス表現
- ⟨TEMPERATURE⟩ 温度表現

- 照応表現
  - ⟨REFORGANIZATION⟩ 組織名参照表現
  - ⟨REFPERSON⟩ 人名参照表現
  - ⟨REFLOCATION⟩ 地名参照表現
  - ⟨REFARTFACT⟩ 固有物参照表現
  - ⟨REFDATE⟩ 日付参照表現
  - ⟨REFTIME⟩ 時刻参照表現
  - ⟨REFMONEY⟩ 金額参照表現
  - ⟨REFPERCENT⟩ 割合参照表現

- 補助表現
  - ⟨PTITLE⟩ 役職・敬称表現

以下、追加したタグについて、簡単に解説する。

3.1 時間間隔表現

固定された時間表現ではなく、単純に時間間隔を表現している場合は、時間間隔表現として抽出する。

<PERIOD> 2時間 </PERIOD>
<PERIOD> 五週間 </PERIOD>
それ以降の <PERIOD> 2 3年間 </PERIOD>
<PERIOD> 1 5分間 </PERIOD>

3.2 電話番号、電子メールアドレス等

電話番号、電子メールアドレス、URL は固有表現として抽出する。

電話は <PHONE> 0 3 − 3 5 0 9 − 8 6 6 1 </PHONE>
<EMAIL> f o o @ s o m e . w h e r e . j p </EMAIL>
<URL> h t t p : ／／w w w . s o m e . w h e r e .
j p </URL> にて公開中

3.3 単位を伴う数値表現

以下のような単位を伴う数字を抽出する。ただし、単位が省略されている場合も、その単位があるものとして抽出する。

なお、点数表現 ⟨POINT⟩ は、厳密に何点と数えられるものに絞って抽出する。具体的には、点数表現候補の部分を「N点」と置き換えて意味が変わらないものを抽出する。例えば、野球のN打点、Nセーブなどの表現は、単純な点数とは別の単位と考えて抽出しない。

<LENGTH> １０メートル </LENGTH>
<SQUARE> １００坪 <SQUARE>
<VOLUME> １０ミリ立方メートル </VOLUME>
<NPEOPLE> ８４人 </NPEOPLE>
<NORGANIZATION> 4社 </NORGANIZATION> で
<NARTIFACTS> ２２本 </NARTIFACTS>
<NARTIFACTS> １０個 </NARTIFACTS>
<AGE> １７歳 </AGE>
<AGE> ２０代 </AGE>
<POINT> ９０点 </POINT>
<SPEED> 時速１００キロ </SPEED>
<FREQUENCY> １００MHｚ </FREQUENCY>
<CSPEED> １００Ｋｂｐｓ </CSPEED>
<RESOLUTION> ７２ｄｐｉ </RESOLUTION>
<VOLT> ＡＣ５Ｖ </VOLT>
<VOLT> ＤＣ＋２４Ｖ </VOLT>
<CURRENT> ２０Ａ </CURRENT>
<CURRENT> ２０アンペア </CURRENT>
<WATT> 5Ｗ </WATT>
<WEIGHT> ２０ｋｇ </WEIGHT>
<PULSE> ２０ｐｐｓ </PULSE>

3.4 照応

抽出対象となっている固有表現、時間表現、数値表現を参照している表現は照応表現として取り出す。役職は人名に伴って使われている場合は、役職名として取り出すが、単独で現れて先の人名を参照している場合は人名に対する照応表現として抽出する。人名の全部または一部が繰り返し使われていても、照応表現とせず、人名として抽出する。照応的表現で時間を表わしている場合は照応時間表現として抽出する。これは、指示代名詞を利用した複合的表現でも同様である。

<REFORGANIZATION> 同社 </REFORGANIZATION> は
<REFPERSON> 彼 <REFPERSON> は
<REFLOCATION> そこ <REFLOCATION> では
<REFARTFACT> それ <REFARTFACT> を
<REFDATE> 同日 </REFDATE> は、
<REFTIME> その時間 </REFTIME> に、

<REFLOCATION> 両国 </REFLOCATION> が領有権を...
<REFMONEY> その料金 </REFMONEY> で、
<REFPERCENT> その割合 </REFPERCENT> で、

<PERSON> 千葉真一 </PERSON><PTITLE> 社長 </PTITLE>
が担当する。．．．<REFPERSON> 社長 </REFPERSON>
は

<PERSON> 千葉真一 </PERSON><PTITLE> 氏 </PTITLE>
が選ばれた ．．．<PERSON> 千葉 </PERSON><PTITLE>
氏 </PTITLE> は

照応関係の抽出を考えた場合、照応関係の単位がIREX の定義と整合しない場合がある。これは、組織名や時間の表現についてできるだけ大きな単位で抽出するという方針に関係している。例えば、

<ORGANIZATION> ○○大学工学部 </ORGANIZATION>

のように組織名をまとめて抽出する。これは組織名に上下関係があるためであり、

<ORGANIZATION> ○○大学 </ORGANIZATION>
<ORGANIZATION> 工学部 </ORGANIZATION>

のようにはしない。ここで、この文のあとに「同大学理学部...」という記述があった場合に照応表現をどうとるかが問題となる。
現在は、

<REFORGANIZATION> 同大学 </REFORGANIZATION> 理学部

としている。もし、

<REFORGANIZATION> 同大学理学部 </REFORGANIZATION>

とすると、異なる組織間での照応を表してしまう。このような、現象は「同年１０月」のような時間表現についても発生する。

3.5 役職名・敬称

敬称、役職名を補助的な表現として、⟨PTITLE⟩として抽出する。ただし、一般的な役職を表している場合は抽出しない。抽出対象となる役職名は、特定の職業や職位を表しているものである。主婦や容疑者など役職名ではない表現は対象としない。また、さらに大きなまとまりで別の固有表現になる場合は、そちらを優先する。役職名が単独で現れて、照応関係を表している場合は、⟨REFPERSON⟩とする。

～ <PTITLE> 氏 </PTITLE>
～両 <PTITLE> 氏 </PTITLE>
～の各 <PTITLE> 氏 </PTITLE>
～ <PTITLE> さん </PTITLE>
～ <PTITLE> 議員 </PTITLE>
～ <PTITLE> 監督 </PTITLE>
～ <PTITLE> 選手 </PTITLE>
～ <PTITLE> 弁護士 </PTITLE>
～ <PTITLE> 医師 </PTITLE>
～ <PTITLE> コーチ </PTITLE>
～ <PTITLE> トレーナー </PTITLE>
～ <PTITLE> 教諭 </PTITLE>
～元 <PTITLE> 首相 </PTITLE>
～ <PTITLE> 会長 </PTITLE><PTITLE> 夫人 </PTITLE>
～ <ORGANIZATION> 事務局 </ORGANIZATION><PTITLE>
長 </PTITLE>
～ <ORGANIZATION> 長野県 </ORGANIZATION><PTITLE>
議 </PTITLE>
～ <PTITLE> 国対委員長 </PTITLE>
～ <ORGANIZATION> 蔵 </ORGANIZATION><PTITLE> 相
</PTITLE>

<ORGANIZATION> ○○知事後援会 </ORGANIZATION>
～ <PTITLE> さん </PTITLE> は野球チームの監督
次の首相は
<REFPERSON> 首相 </REFPERSON> の指示で
三役
～（主婦）
～候補
～容疑者
～ <ORGANIZATION> 専売局 </ORGANIZATION> 職員が

4 トランスデューサによる固有表現抽出

この節では、トランスデューサにより固有表現を抽出する方法について述べる。

4.1 トランスデューサ

有限状態オートマトン (FSA) は、状態遷移を表す矢印のラベルに有限アルファベットの１つの要素が入力として付与されているが、有限状態トランスデューサ (FST) では、入力と出力を表す２つのシンボルがラベルに付与されている。トランスデューサの詳しい定義や性質については、文献 [3] に譲るが、直観的にいえば、トランスデューサとは、入力列に対して、状態遷移をする毎に出力を出すような有限状態オートマトンである。

4.2 トランスデューサの応用

我々が開発中の固有表現抽出ツール ProCreator は、ある正規表現を満たすパターンの入力に対して、置

換を行なった結果を出力するという簡単なトランスデューサの集合として構成されている。

入力となるアルファベット Σ は、形態素の集合 M、品詞の集合 P、意味マーカの集合 S、固有表現タグの集合 T とすると、

$$\Sigma = (M \times P \times S^*) \cup (M \times T) \cup \{\varepsilon\}$$

であり、アルファベットの要素を $< m,p,s >$ または $< m,t >$ (但し$m \in M, p \in P, s \in S^*, t \in T$) と書く。また、ラベルに書くアルファベットの要素の略記として、ワイルドカードを意味する * を導入する。例えば、形態素や意味マーカに関わらず、品詞 p により状態遷移する場合には $< *,p,* >$ とかく。

簡単なトランスデューサの集合として、固有表現抽出ツールを構成することにより、各タグの表現抽出のための保守性の向上、およびタグ自体の追加、削除の効率化が期待できる。

4.3 トランスデューサの実装

最終的には高速に動作するトランスデューサとして実装する予定であるが、現在、評価のためにトランスデューサを論理的な表現として採用するだけで、実際の実装は Perl の置換により実現している。

現在、約３００の置換が人手によって作成されている。実験による評価はまだ行なわれていないが、IREX のタグの範囲については、４～５月に行なわれる IREX の固有表現抽出に関するコンテストに参加し、評価をする予定である。

以下、置換によりタグを付与する２つの例を示す。

（例１）

役職名の前に人名の可能性のある固有名詞があれば、人名タグをつける。

```
<千葉,固有名詞,{人間,場所}>
<総裁,<PTITLE>>
```

という入力列に対して、

```
<千葉,<PERSON>>
<総裁,<PTITLE>>
```

という出力結果を出力する。

（例２）

```
<千葉,固有名詞,{人間,場所}>
<真一,固有名詞,{人間}>
```

という入力に対して、

```
<千葉真一,<PERSON>>
```

という出力を出す。

例１、例２ともに最終的には、それぞれ

```
<PERSON>千葉</PERSON><PTITLE>総裁</PTITLE>
<PERSON>千葉真一</PERSON>
```

のようにタグで前後を括った形式に変換して出力する。

次に、このような置換を Perl により実装する方法について述べる。基本的には、置換 s を使う。$line に改行で区切られた一文の形態素／意味解析結果が代入されているとする。なお、各行は空白で区切られた形態素、品詞名、意味カテゴリ集合の列とする。

```
$morph =~ s/パターン/置き換え対象/gm;
```

Perl のユーザには自明であるが、オプション g により、一文にマッチするパターンが複数あった場合に、全体に対して置換が行なわれる。また、オプション m により、文字列を複数行として扱うことが可能なる。

このような設定により、複数行に渡る形態素情報のパターンに対するマッチングとそれに対する置換を１つの非常に簡単で高速なコマンドで実現できる。例えば、例１のタグの付与は、次のような置換により実現できる。

```
$morph =~ s/^(.*?) 固有名詞 .*? 人間.*?\n^(.*?)
<PTITLE>\n/$1 <PERSON>\n$2 <PTITLE>\n/gm;
```

5 おわりに

本稿では、まず、情報抽出システムでの利用を考え、日本語固有表現抽出のためのタグの拡張について述べた。次に、これらのタグを出力する日本語固有表現抽出プログラムをトランスデューサにより実現する方法について述べた。今後、現在作成中の正解タグ付きの３０００記事を用いて、固有表現抽出プログラムの性能を測定していく予定である。

参考文献

[1] 関根 聡, 江里口善生: 固有表現の定義の困難さ – IREX における NE 定義の苦労話 –, 言語処理学会第５回年次大会, 1999.

[2] 廣田啓一, 佐々木裕, 加藤恒昭: オントロジ主導による情報抽出手法の提案, 言語処理学会第５回年次大会, 1999.

[3] E. Roche and Y. Schabes: Finite-state Language Processing, MIT Press,1997.

[4] IREX ホームページ. (http://cs.nyu.edu/cs/projects/proteus/irex/)