固有表現の定義の困難さ
– IREXにおけるNE定義の事例から –

関根聡
New York University
sekie@cs.nyu.edu

江里口善生
(株) NTTデータ 情報科学研究所
eriguchi@lit.rd.nttdata.co.jp

1 イントロダクション

組織名、人名、地名、商品名のような固有名や、時間表現、数値表現は新聞記事などの文章を理解するためにキーになる要素である。それらを自動的に認識し、分別することは自然言語処理の第一段階として非常に有効であると考えられる。その技術を目的としたコンテストも企画されている (IREX) [1]。しかしながら、固有表現を定義すること自体、非常に難しい問題を含んでいる。(この論文では、固有名、時間表現、数値表現をまとめて固有表現と呼ぶ。) この問題には、意味やプラグマティックスの問題も含まれ、人によっても異なる判断が得られる場合もある。

本論文では、IREXにおける固有表現の定義を簡単に紹介し、そこで特に問題になった点について詳しく述べ、固有表現について考察を試みる。誌面の都合上、定義などは簡略化してあるが正式な定義やメイリングリストにおける議論などはIREXのホームページ [1] にあるので参照されたい。

2 MUC, MET, IREXの歴史と固有表現

固有表現 (Named Entity) 抽出という課題は、米国の情報抽出のコンテストであるMessage Understanding Conference (MUC) の1996年に行なわれた第6回目のコンテストで導入された [2] [3]。背景には、イベント情報などの情報抽出という課題において、抽出するための知識 (辞書、文型パターン等) を対象タスク (企業合併、人事異動など) によって作成しなおさなければいけないという問題点があった。この問題に対処する為、固有表現抽出という課題が対象タスクに依らないサブコンポーネントとしての位置付けで作られた。MUCは主に英語を対象にしているが、それから派生したMultilingual Entity Task (MET) は、日本語をはじめ、いくつかの言語を対象とした固有表現抽出のコンテストで、MUC6,7と並行してMET-1,2が開かれた [4]。この流れを汲んで、日本でも独自のコンテストを開く気運が広まり、1999年4,5月に行なわれるInformation Retrieval and Extraction Exercise (IREX) の1課題として、固有表現抽出の課題が設定された。MUC, MET, IREXにおいて、固有表現の定義は参加者の議論や、コンテストの意図で変ってきているが、新しいコンテストにおける定義は過去の定義を踏台にしており、この3つのコンテストはひとつの流れとして捉えられる。この論文ではIREXにおける固有表現定義に基いて議論をするが、より興味のある読者にはMUC, METを参照されることを薦める。

3 固有表現の種類

IREXでは、固有表現の種類として以下の8種類を用意した。MUC,METでは7種類であるが、商品名等の固有物名 (ARTIFACT) を追加した。また、役職名も導入しようという意見もあったが、今回は見合わせた。他にも、イベント、固有の事の名前などの新たな種類の固有表現も考えられる。

固有名詞的表現

- 組織名 (ORG)
  複数の人間で構成され、共通の目的を持った組織等の名称である。株式会社等の会社、固有の政府組織、学校、軍、スポーツチーム、国際組織、労働組合、工場、ホテル、空港、病院、教会やなんらかの目的を持ったグループ等もその対象が組織としての意味で使われている文脈においては組織名とする。

- 人名 (PERSON)
  固有の人を指す名前。

- 地名 (LOC)
  固有の場所を指す名称。大陸、国名、地域名、都市名、地方名、県名、町名、村名、道路名、住所、駅名、線路名、モニュメント、海洋名、湾、運河、川名、池名、湖名、島、公園、山、砂漠の名前などを含む。

- 固有物名 (ARTIFACT)
  人間の活動によって作られた具体物、抽象物を含む固有の物の名称。

時間表現

- 日付表現 (DATE)、時刻表現 (TIME)
  時間表現では、絶対的な表現 (1999 年 2 月 1 日など) や，基点が明確であり絶対的な時間が分るような相対的な表現 (昨日など) を抽出する。日付表現は、その単位が 24 時間以上である物を指し、時刻表現は、その単位が 24 時間以下であるものを指す。

数値表現

- 金額表現 (MONEY)
  金額を表わす表現。
- 割合表現 (PERCENT)
  割合を表わす表現。

また、IREX では、定義上どの種類の固有表現となるか判断が難しい物や固有表現の認定が難しい物などを OPTIONAL として、コンテストの際には評価対象から除いた。データ中ではこれらの固有表現には SGML のタグが付与されており、データは例えば以下のようになっている。(それぞれ、上記の説明で括弧内にあるタグを該当文字列の先頭と最後に振っている。)

<DATE> 今年 </DATE> は <OPTIONAL POS=LOC,ORG> 関空 </OPTIONAL> 効果で <LOC> 大阪 </LOC> への観光客は <DATE> 前年 </DATE> の <PERCENT> 3 割 </PERCENT> 増しと予想されている。

4 固有表現の定義と問題点

前述のように固有表現というものを定義し、その種類を定めた。これらの定義に基づけば、問題なく文章中の固有表現を同定できそうであるが、実際にデータをみてみると一筋縄ではいかない色々な問題が発見された。以下に、その内の重要なものについて紹介し、固有表現というものについて考察してみる。

本論文では、固有表現定義における問題点を具体的に挙げ、そこには本質的な難しさが存在するということを主張する。実際の IREX での定義は極力触れないので、興味のある方には IREX ホームページを参照していただきたい。

4.1 表記か意味か

「固有の物の表現」というと、固有名詞が連想される。しかし、固有名詞だけが固有の物の表現ではない。例えば、「ニューヨーク大学」は固有名詞であり、「大学」は普通名詞である。単語だけで考えた場合は、これに疑問の余地はないが、実際の文章では照応等で「大学」という表現を利用して、「ニューヨーク大学」を指しているという場合がある。(例文「ニューヨーク大学の紹介。大学はマンハッタンの中心部にあり、…」) したがって、意味的には、または文脈に基づけば、この「大学」も「固有の物の表現」であると言える。つまり、表記だけで固有表現を定義するとすると、「大学」は固有表現にはならないが、意味に基づくと「大学」も固有表現にもなることがあると言える。固有表現を定義する際に、「表記に基づくか」と「意味に基づくか」という基準は連続的であり、それらの中間的な表現も存在する。例えば、普通名詞と思われる「国会」がそうである。「国会」は、日本という限定を置けば、ひとつしか存在しない。したがって、普通名詞といえども、文脈を使用しないでも、ほぼ固有の物の表現と言える。より困難な例は、「県議会」「都議会」である。日本には県は 43 あるが、都はひとつしかない。「県議会」は、文脈がないと固有のものかどうかは判断できないが、「都議会」は、文脈なしで (世界知識と表記だけで) 固有のものと判断できる。ここまでくると、文脈はもとより世界知識といった物が必要となってきており、客観的な基準を作成しようとする際に非常に重要な問題を投げかける。つまり、世界知識は個人によって異なるため、客観的な定義は不可能だという結論になる。

4.2 表記か意味か：その 2

固有名の認定の場面だけではなく、固有名の種類を決定する場合にも、表記によるか意味によるかという問題が生じる。

<LOC> アメリカ </LOC> の圧力，<LOC> 永田町 </LOC> の決断

これらの表現は、表記上は地名であるが、その意味はアメリカ政府や日本政府といった組織である。特に、新聞記事では国名などの地名がこのような形で使用されることは多い。文章を理解するためのキーになる要素として固有名をとらえる場合は、意味を基準に判定する方がよいと思われる。しかし、意味には、読者個人的な概念背景や文脈も影響するため、全ての人が同じ判定を下せるような定義を作るのは難しい。

4.3 表記か意味か：その 3

IREX の定義では、時間表現は、絶対的な時間軸における時が分るような時間表現としている。固有名詞の場合と違い、普通名詞のみで書かれている場合もあるため、

表記だけで判断するというのは困難であり、もともと意味に基づいて定義をする必要がでてくる。例えば、「夏」や「年」という文字があっても、それがある特定の夏や年を指している場合には、それを時間表現としなければいけないが、そうでない場合も存在する。「夏」には少なくとも、「ある特定の夏(例：１９９９年夏)」という意味と、「クラスとしての夏(例：夏は暑いものだ)」という意味がある。後者の「夏」は特定の時間を指すものではないため、時間表現ではない。したがって、このような表現では表記ではなく、意味のレベルで分ける必要があった。

4.4 固有名から普通名詞への変化

新しい商品が開発された時に付けられる名前は固有名であるというのは、ある程度曖昧性なく判断できる。しかし、その商品が広く売れて、その商品名が一般名詞的に使われることがよくある。「ウオークマン」は典型的な例である(その他、セスナ、エレクトーン、ラジコン、セロテープ、セメダイン等ある)。本来はソニーが作った特定の商品の名前であるが、一般的には、製造メーカーにこだわらず、スピーカーの付かない小型のカセット再成機やラジオなどがこの名前で呼ばれる。英語でも「bath(風呂)」のように、現在は完全な普通名詞として使用されている物も、語源まで辿ると実は固有名詞である場合もある。このように、固有名詞と普通名詞には、時間軸に沿った連続性が考えられる。しかし、固有表現を定義する際には、このような語源まで考えるのは意味があるとは思えない。また、現実的に語源の知識は調べない限り、知り得ないという場合が多いので、ユニークに定義するのは困難である。

4.5 商品名とクラス名

辞書によると、固有名詞とは「同じクラスにある別々の物を指すためのラベル」とある。しかし、特に商品の名前などに顕著であるが、クラス自体も固有的な性格を持つ物であったり、クラスが階層的になっている場合にはどの部分を持って固有表現とすればいいか判断が非常に困難な場合がある。例えば、「トヨタ車」「カローラ」「カローラセダン」「カローラセダンSE-Saloon」または、「各車についている車体番号」のどこまでを商品名としてよいかを明確に定義するのは困難な問題である。

4.6 部分表現の扱い

英語の場合と異なり、日本語では単語の間にスペースがないので、表現の範囲も曖昧になっている。例えば、「来日」という表現において、「日」というのは「日本」のことであるので、固有表現としたいという立場と、「来日」はひとつの単語と見做せるので、その内部の表現は固有表現としたくないという立場がある。後者は、辞書にある表現なら単語とする、というような人工的な規則を作るとある程度は納得のできる定義になるが、例えば、「来日」が辞書にあったとすると、その同類の表現で「在エジプト」の場合の「エジプト」はどうずべきかといった問題がでてくる。

逆に部分表現でも固有の物を指す表現はすべて固有表現とすれば問題がないかというと反例も出てくる。「オーデコロン」の「コロン」は地名に由来するが、この「コロン」を固有表現として抽出するのに意味があるか疑問である。

同様に、「日本語」「英語」や「フランス料理」の部分を地名として説明するのは、語源まで遡ることに類似しており難しい。

4.7 照応、省略形

今回のコンテストでは、照応表現は抽出しないものとした。(これは、MUCにおいて照応が別の課題として設定されていることに少なからず関係している。)代名詞等を固有表現とすると、最初に提起した「表記と意味」の議論において、意味に基づく固有表現定義の極端な例となる。また、日本語の照応にはゼロ代名詞など難しい問題が存在することも問題を複雑にしている。

新聞記事において、会社名などの組織名が正式な名前で出ているのは、せいぜい最初の１回目だけということが多い。また、「北朝鮮」は「朝鮮民主主義人民共和国」が正式名であるが新聞記事においてこの正式名はあまり使用されない。また、文脈が進んでいくにつれ「北」だけで北朝鮮を指す場合もある。これらは、単なる省略形として判断すべきなのか照応なのか判断が難しいところである。

4.8 抽象物、物と事

固有物名には、抽象的な物も含むとした。しかし、抽象的な物は、人間が目で見ることができないという理由により、客観的な定義はなかなか難しい。抽象的な物を説明する為に、たとえば、著作権、知的所有権が主張可能であるような物を固有物として定義するという方法がある。それによれば、作品名、出版物、成果物はかなりカバーできる。しかし、法律名、法案名、条約名、学説名や俗説名、制度、税の名前のようなものは、どのように判断していいか難しい。

関連する議論としては、「物」と「事」の境が曖昧であるという問題もある。名詞には大雑把に「物」と「事」があるとされている。「物」は静的に存在する物であり、「事」は動的な叙述内容の表現である。今回のIREXでは、「固有な事」というのは判断が難しいため、課題と

して取り上げるのを諦めた。しかし、商品名を始うとする固有な物を定義する際には物と事の境が問題になってくる。例えば、以下のリストでは、どれが「物」であり、どれが「事」であるかという判断は人によって異なると思われる。

本の題名、ビデオの題名、テレビ番組名、講演名、会議名、集会名、戦争名

サービス名

これも抽象的な物の名称であるが、商品名において商品がサービスである場合に、そのサービスの名前が明確でない場合も多い。「雪見だいふく」という商品の場合には、その物が具体物であり、手に取って見せることができるのも手伝って、それを固有物名と言うのは難しくないように思える。しかし、6：00東京発の「のぞみ1号指定席券：7号車：座席番号23A」を買った場合に、そのサービスの名前(または商品名)が何であるかを決めるのは容易でない。

ニックネーム

ここで議論するニックネーム(愛称、別称)に関する問題は時間表現に限った物ではないが、特に時間表現において顕著に見られるので、それを題材として説明する。

例えば、1999年2月1日というのは、幅広い人が理解できる絶対的な表現と言える(もちろん、本質的には限定的、相対的である)。また、江戸時代という時間に関するニックネームも日本の一般的教育を受けた人間なら特定の時間表現として認識できる。しかし、ニックネームは慣用的な表現であり、その有効範囲は幅広いとは限らない。例えば、「母の日」というニックネームの類似物として「義母の日」という日が存在する。この由来は著者には不明であるが、これを一般的な時間表現であるとするには抵抗がある。また、より限定された例では、「燃えないゴミの日」といった物もある。これらは、明らかに限定的であるが、時間表現には代りない。どの程度、慣用的であるか、流布しているは主観的に決定せざるを得ない。

5 考察

固有表現の定義はこれまでの例でも分るように曖昧性を含み難しい問題である。これらの問題は実際にデータを作成する過程を通して広く見付けられた。個人によっても揺れるし、文脈等に強く依存する。最終的には主観的に決めざるを得ない部分がある。したがって、明らかにIREXの定義は単なるひとつの切口でしかない。

抽出された固有表現の使用目的を限定することによって定義の曖昧性をかなり減らせると考えられる。例えば、日本の要人の動きを知りたいという目的であれば、「天皇の訪米」における「天皇」は人名として扱うべきである(天皇は役職名とも考えられるので、IREXの定義ではOPTIONAL)。ただし、定義を明確にしたいがために単に目的を絞っていいかというと、技術の汎用性や応用分野の幅を維持するという観点とのバランスが必要であると考えられる。また、数多くの団体が参加するコンテストにおいて、あまりに限定された課題や目的を設定することが妥当かどうかという問題がある。つまり、固有表現抽出を独立したコンテストにするという立場においてはなるべく汎用的な定義を作成することは意味があるものと考えている。

ただし、MUCのように、対象を限定した情報抽出の一部という位置付けとして、特定の目的や分野での固有表現抽出という課題も考えられるし、IREXではその種類の課題も作成している。

6 謝辞

IREX参加者ならびに固有表現定義の議論に参加してくださった方に感謝する。特に、沖電気の福本氏、富士通研究所の落谷氏、松下電器の野口氏、ＮＥＣの竹元氏、東京大学の野畑氏、通総研の井佐原氏、内元氏には、貴重な意見をいただいた。

参考文献

[1] IREX homepage: http://cs.nyu.edu/cs/projects/proteus/irex/

[2] Message Understanding Evaluation and Conference *ARPA, Morgan Kaufmann Publishers* You can place your order through SAIC's MUC homepage (below)

[3] Ralph Grishman, Beth Sundheim: "Message Understanding Conference - 6: A Brief History" *COLING-96*

[4] MUC homepage: http://www.muc.saic.com/