機械翻訳における固有名詞処理

島津美和子　　吉村 裕美子
(株) 東芝　研究開発センター

1 はじめに

近年、インターネット・WWWなどへの関心の高まりから、英日機械翻訳システムへの注目も後を追うように急激な高まりを見せている。これは、従来より産業翻訳で求められてきた訳文書を作成するための翻訳システムではなく、英文を読むための翻訳機能を求めるものである。かつ、ユーザ層も翻訳を業務とする専門家から、特に言語知識を持たない一般人へと変化している。

ウェブ文書の特徴の1つに固有名詞の頻出があげられる。固有名詞は原文書の意味を読者に伝える上でキーワードの1つになりえるため、それが適切に翻訳されるか否かで、読者に与える印象に大きな差をもたらす。従来の産業翻訳では、新規・未知の固有名詞に対してはユーザが辞書に登録することで対処可能であったためこのことは問題とならなかったが、ウェブ用の翻訳では、ユーザが言語の専門知識をもたないことや、また単に手間がかかるという理由でユーザ登録が期待できないため、新たに問題として顕在化してきた。

情報検索やメッセージ理解などの分野でも固有名詞の正しい認識は重要な問題であり、研究も盛んに行われている（[Chen & Lee, 1996]、[Gallippi, 1996]、[Strzalkowski & Wang, 1996]、[若尾, 1996]）。しかし、これらは幅広く一般的に固有名詞を認識することが目的であるため、翻訳に関連した特有な問題は残っている。

本稿では、上記特徴に対応して、翻訳品質を高めるために構築した固有名詞処理の機能とその評価結果について報告する。

2 固有名詞の翻訳における問題

新規（辞書に未登録）の固有名詞が用いられた場合に一番問題となるのは、それが強調のために大文字化されただけの一般語なのか（これは小文字化した語として辞書引きし、訳出されなくてはならない）、あるいは、人名、団体名などの固有名詞であるかが区別しにくい点である。固有名詞であれば一般には原語つづりのまま訳文中に出力するという対応策がとれる。しかし、構成語がその固有名詞の意味を示すものである場合には、小文字化した語に分解し訳語を合成して訳出するほうが望ましい。（例1）はその例である。逆に、（例2）の「Apple」、「Gene Glazer」は「apple」，「gene」，「glazer」として翻訳すべきでない例である。特に、人名、会社名、地名などがこの類に属する。「Gene Glazer」は「Mr.」や「said」の存在などから人名であることが認識できる。そのため、人間なら、後続文中に（例3）のように単独で「Glazer」が用いられても人名であると理解できる。（例2）の「Apple」、（例4）の「Prudential」もまさにそのケースであり、先行文に「Apple Computer」という句が存在すれば、後続文で「Apple」と単独で登場しても読者はそれを果物名だとは思わないし、先行文に「Prudential Insurance Co.」という語があれば、「Prudential」を形容詞句であるとは思わない。これは、機械翻訳では訳語の不適切さばかりでなく構文解析の失敗をも招くことがある。例えば、（例4）でprudentialが形容詞であることから翻訳システムが「Prudential」を形容詞句と認識すると構文解析の失敗の原因となる。

（例1）
E：This is Beijing Automation Technology Research Institute.
J：これは北京オートメーション技術研究所である。

（例2）
E："Apple has to do something," said Mr. Gene Glazer.
J：「りんごは何かを行わなければならない。」と遺伝子つや付け工が言った。

（例3）
E：In the joint briefing Tuesday, Glazer unveiled ...

（例4）
E：Prudential says it will ...

Processing of proper nouns in machine translation
Miwako SHIMAZU and Yumiko YOSHIMURA
Toshiba Corporation

一方、「Apple Computer」、「Gene Glazer」は辞書に登録されているが「Apple」、「Glazer」は登録されていない場合でも、後続文の「Apple」,「Glazer」の訳語は、「Apple Computer」、「Gene Glazer」の訳語定義から「Apple」,「Glazer」に対応する部分を自動抽出して訳出できることが望ましい。「Apple Computer」の訳語と「Apple」単独の時の訳語が異なることは、一般ユーザが訳文を読む上で混乱につながる。

以上のように、固有名詞を正しく認識・訳出できないと、構文解析の失敗や、原文の意味を損なった訳文を生成してしまうことがある。そのため、人名、会社名、地名などは文脈から固有名詞であるかどうかを推定するとともに、後続の文の翻訳へもそれを伝搬させる機能が必要である。固有名詞を正しく固有名詞と認識することは、その名称の訳出を改善するだけでなく、それが人名、団体・組織名、地名と認識されることで、それと共起する語の訳し分けを改善させる可能性があるという点でも重要である。

3 固有名詞の推定

文中にある未登録固有名詞の推定は、(1) 一文中の固有名詞が存在することを示すキーとなる語（句）を用いる処理と、(2) 前出文中の固有名詞情報をもとに後続文中の固有名詞（特に部分語）を推定する処理からなる。本節では (1) について述べ、次節で (2) について述べる。

(1) は、まず大文字で始まる語が２語以上連続するものに関し、固有名詞であるかどうかを推定し、固有名詞と判定された範囲にある語に対しては、もともと固有名詞の語があれば固有名詞を優先し、固有名詞がない場合、すなわち、動詞や形容詞などしかない場合には、辞書引き結果として固有名詞の候補を追加する処理である。下の（例５）の場合、固有名詞の推定が行われないと、「United」 の辞書引き結果としては過去・過去分詞しか存在せず、構文解析が失敗するが、「Inc.」をキーにして前接の「UtiliCorp United」が会社名の構成要素であると判定されれば、辞書引き結果として固有名詞が追加され、解析の失敗も回避される。（但し、品詞は固有名詞と認識されても、訳語の作成はできないため、以下のように英語のつづりのまま訳出される。）

（例５）
E: This is UtiliCorp United Inc.
J: これは UtiliCorp United 社である。

判定する固有名詞の種類は団体名、人名、地名であり、判定に用いるキーは以下の通りである。（合計約８０種類）

- 団体名
  - 最後の語が会社等団体を示すキーワードである。
    (e.g.「Corp.」「Ltd.」)
- 人名
  - 先頭の語がタイトルを示す語である。
    (e.g.「Mr.」「Dr.」)
  - 直前の語が役職、職業を示す語である。
    (e.g.「spokesman」「chairman」)
  - 人名であることを強く示す語句が隣接する。
    (e.g.「, who/whose」など)
- 地名
  - 最後の語が地名を示すキーワードである。
    (e.g.「River」「Island」)

4 文脈を利用した固有名詞の部分語の処理

通常、人名や会社名は、初めに正式名称（すなわちフルスペル）で現れ、その後その一部分だけが使われる。先に示した（例２）の「Apple」、（例３）の「Glazer」、（例４）の「Prudential」がこれに該当する。そこで、推定されたものも含め、生起した複数語からなる固有名詞のキー構成要素を記録し、後の文の処理で参照できるようにした。これにより、先行文脈に会社名として登録されている「Apple Computer」があるが「Apple」単独では登録されていない場合にも、「Apple」が会社名である解釈を優先する。かつ、「Apple Computer」の訳語定義の語の構成を解析し、辞書中の訳語定義が「アップル・コンピュータ」であるなら、「Apple」の訳語として、対応する「アップル」を生成する。

この訳語の学習機能は人名の訳出にも有効である。人名は通常、アルファベット発音をカタカナ化した文字列で書かれる。人名は同じスペルでも発音が異なることがあったり、同じ発音でも外国人名が日本に紹介される際に異なったカタカナにが当てられ定着することがある。例えば、「Hepburn」には、「ヘボン」と「ヘップバーン」という２種のカタカナ訳語が存在する。しかし、「ヘップバーン」の方が原語の発音により近く、西洋の苗字として一般的であるのに対し、「ヘボン」は「James Hepburn」という特定の人を指示する名前として使われる。しかし上記学習機能がないと、「James Hepburn」が辞書に「ジェームズ・ヘボン」と登録されていても、「Hepburn」の訳語定義の第一候補が「ヘップバーン」なら、単独で現われた「Hepburn」に対し、

「ヘップバーン」が生成されてしまい、「James Hepburn」の「Hepburn」と「Hepburn」単独がユーザには同一のものを示す名詞とは映らない。

（例4）の事例については、「Prudential Insurance Co.」が先行文にあれば、それ自体が会社名として登録されていなくても、前節で述べた「Co.」をキーにした推定が行われ、「Prudential」が会社名であるという解釈を優先する。この場合は、「Prudential Insurance Co.」も「Prudential」も共に原語スペルのまま訳文に生成する。「Glazer」も同様である。

逆に、先行文脈で小文字の「apple」「prudential」はあるが大文字化した語はない場合には、小文字化した語から訳語合成をできるように、一般語として使われた語も記録しておき、後続文の処理中に参照できるようにした。

5 評価

本機能の効果を評価するために、実際のウェブ文書に対して、本機能を導入した場合としない場合の訳文の変化を調べた。文書は、種々の話題に及び固有名詞の出現の多いニュースページからとった。その種類、数、および実験の環境は以下のとおりである。

- 対象文書
  - ニュースの種類：
    - トップニュース
    - ビジネス
    - 娯楽
    - スポーツ
  - 総語数：4410語
    - 固有名詞数：630語
    - 他内容語：1959語
- 辞書の語彙数
  - 標準語辞書：194099語
  - 固有名詞辞書：33323語

また、タイトル文では一般に、本文に先立ち、固有名詞の一部が用いられるため、従来の翻訳方式では、本文前に翻訳するタイトル文の固有名詞処理に不具合がでやすかった。そこで、今回は、本文の翻訳の後に本文の内容を利用してタイトル文を翻訳するように処理を変更し、タイトル文の翻訳にも影響が現われることをねらった。

上記環境で訳質の変化を比較した結果、差として現われた訳語数と改善率を表1、2に示す。

改善語の種類	改善数	悪化数
固有名詞	24	4
他の内容語	2	0

表 1: 訳語の変化

導入前の不適切訳語数	（改善数－悪化数）	改善率
96	20	20.8%

表 2: 固有名詞訳語の改善率

本機能の導入により全固有名詞中の不適切訳語の19.8%に改善が見られた。中でも、前出したフルネーム人名・団体名の一部が後出するケースにおける品詞・構文的な誤解釈や訳語誤りの回避に対する寄与事例が多かった。また、固有名詞以外の語の訳し分けが改善された事例が2件あった。

4件の悪化事例のうち2件は固有名詞推定範囲の誤りによる。

（例6）は、3つの人名が並列されているが、最初の「German Manuel Reuter」にだけ不具合が出ている。これは「by」が隣接していることが人名推定の要因となっているためである。「German」「American」などの地名から派生した特種形容詞は人名要素の先頭語になりにくいことを判定仕様に追加すればよい。

（例7）は、「Hill」が地名のheadになりやすいということをキーに「Briton Damon Hill」が地名と推定されたために「Briton」が不訳となったケースである。「Hill」は現辞書に、地名要素のほかに人名要素としても登録されている。これは、「Briton」のようにある種の「人」を意味する語が先頭にある場合には地名推定の対象から排除するようにすれば回避可能である。

（例6）

E: The leading Porsche is driven by German Manuel Reuter, American Davey Jones and Austrian Alexander Wurz.

J(旧): 主要なPorscheは、ドイツのマヌエル・ロイター、アメリカのデーヴィ・ジョーンズおよびオーストリアのアレグサンダーWurzによって運転される。

J(新): 主要なPorscheは、Germanマヌエル・ロイター、アメリカのデーヴィ・ジョーンズおよびオーストリアのアレグサンダーWurzによって運

転される。

(例7)
E: In Formula One, Briton Damon Hill has the pole for Sunday's Canadian Grand Prix in Montreal.
J(旧): F－1では、英国人ダモン・ヒルが、...
J(新): F－1では、Britonダモン・ヒルが、...

残る悪化2件は、固有名詞ではあっても、訳出には小文字見出しに対する訳語を当てたほうがよいという事例である。

(例7)から明白なように、小文字見出しの訳語がアルファベット発音をカタカナ化したようなものなら、いかなる固有名詞と推定されても訳出したほうがよい。訳語がカタカナかどうかは文字種で自動判定できるが、かならずしもカタカナの訳語が常に原語の発音をもとにしたものではないので、厳密には、訳語に弁別素性を付けてやらないと解決できない。

一方、(例8)は、前出の(例1)と同様、小文字見出しの訳語を訳出すべきか否かで正否の判定ができずに残る部類である。「Theater/Theatre」は「Neil Simon Theater」「Fox Theater」のように前に人名、地名、団体名など伴うことが多い。また、「Colonial Theatre」「Majestic Theatre」などを小文字化した訳語の合成から「植民地の劇場」「雄大な劇場」と訳出すると固有名詞と認識できなくなる。それを回避するために「Theater/Theatre」,「Street」のように小文字では普通名詞だが大文字化して固有名詞のheadとなりやすいものについてはその名詞の前に来る修飾句を原語のまま出力したほうが望ましい固有名詞の推定キーに含めているが、(例8)の「Actors」ように小文字語から実体を示す語に訳出することにより意味理解を助けるケースと上記の「Fox」、「Colonial」、「Majestic」のように原語スペルのままのほうが固有名詞らしさを保てる点で望ましいケースととの識別は難しい。このようなものに対しては、大文字名詞句の後に原語スペルを併記することでも、ユーザの理解を助けることができる。補助的ではあるがユーザ支援としては十分に機能する。

(例7)
E: They were found in Cocoa Beach along a stretch of oceanfront condominiums and were taken to the Kennedy Space Center for examination.
J(旧): ...ココア海岸で見つけられ、...
J(新): ...Cocoa海岸で見つけられ、...

(例8)
E: Scott co-stars with Charles Durning in the National Actors Theater revival of the Jerome Lawrence/Robert E. Lee drama.
J(旧): ...のドラマの全国俳優劇場リバイバルで、...
J(新): ...のドラマの全国Actors劇場リバイバルで、...

(原語スペル併記の訳出)
(例7)' ...ココア海岸(Cocoa Coast)で見つけられ、....
(例8)' ...のドラマの全国Ａｃｔｏｒｓ劇場(National Actors Theater)リバイバルで、...

6 おわりに

ウェブ文書の翻訳品質向上のために、高頻度が特徴的である固有名詞の訳出改善のアプローチを提案し、それに対する評価を行い、その効果を数値的に確認することができた。今後は、コーパスを材料に仕様の改良を行うと共に、推定知識の収集を行っていく。

参考文献

[Chen & Lee, 1996] Chen, Hsin-Hsi and Jen-Chang Lee. 1996. Identification and Classification of Proper Nouns in Chinese Texts, In *Proceedings of COLING-96.*

[Gallippi, 1996] Gallippi, Anthony F. 1996. Learning to Recognize Names Across Languages, In *Proceedings of COLING-96.*

[Strzalkowski & Wang, 1996] Strzalkowski, Tomek and Jin Wang. 1996. A Self-Learning Universal Concept Spotter, In *Proceedings of COLING-96.*

[若尾, 1996] 若尾孝博 1996. 「英語新聞記事からの固有名詞自動抽出技術」 自然言語処理 115-9. pp. 59-73.