2F－2

英日機械翻訳における固有名詞処理

加藤　直人*　浦谷　則好*　相沢　輝昭*　中瀬　純夫**
*NHK放送技術研究所　**(株)カテナ・リソース研究所

1．はじめに

NHKでは1989年8月から衛星放送の番組「ワールドニュース」の中で、英日機械翻訳システムの試用を開始した。[1] しかし、翻訳品質はまだ十分でなく、翻訳精度向上のためにニュース用語の充実と、ニュース独特の表現への文法の対応を図っていく必要がある。このため英語ニュースデータベースの構築を進めている。

本稿では、この英語ニュースデータベースを用い、ニュースにおいて重要な役職付き人名等の固有名詞を認定、翻訳する方法を考察したので報告する。

2．英語ニュースデータベース

本英語ニュースデータベースは

1）ニュース文の検索
2）ニュース文からの専門語の抽出
3）ニュース文からの文体の抽出
4）辞書・文法の評価

を目的としている。データベースの整備に先だって、「ワールドニュース」の中で実際に使用された文（8月、9月分）とAP電（1989年5月、6月分：要約記事、改作記事等の重複ニュースを除く約209万語）を調査した結果、次のことが判明した。

1）異なり語数（屈折形、大文字形を別に数えたもの）は約57，000語である。語の総数に対する異なり語数の変化を図1に示す。

語数が多くなるにつれ、新規に出現する異なり語中に占める固有名詞の割合が増加（異なり語5千語付近では30%、5万語付近では50%）することが認められた。

2）人名、機関名、国名、地名等の固有名詞が頻出する。

3）広範な用語が使われ、しかも分野による用語分類が難しい。

4）言動、思考に関連する動詞が多用される。say はbe動詞に次いで使われ、have動詞より頻度が多い。call, report, talk, ask, think, want, feelの使用も多い。

5）数字表現の出現頻度が高く（頻度7位）、しかも機械翻訳が困難な複雑なものが多い。

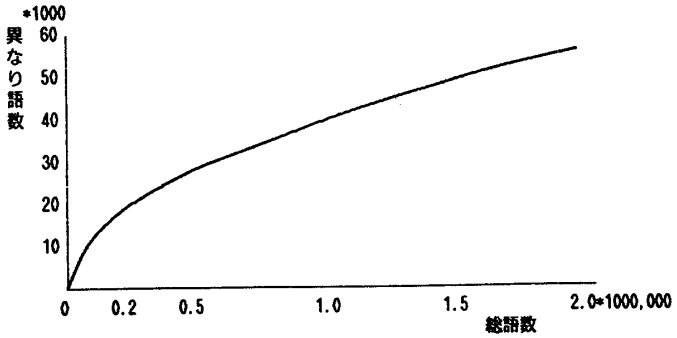

図1：総語数に対する異なり語数の推移

3．固有名詞の解析

2．の結果からわかるようにニュース文には固有名詞が多い。固有名詞の中でも人名、機関名はそれだけ単独で出現することは少なく、前後の単語と共に複合語となる。例えば人名の場合、U.S. President George Bushのように固有名詞 George Bush は役職名 U.S. Presidentと共に複合語となる。このような複合語となる固有名詞をそれぞれ独立した単語として機械翻訳を行なうと、構文解析が複雑となるという欠点が生じる。

また、固有名詞は未知語となることが多い。しかし未知語であっても、前後の単語と共に複合語として扱えば本質的に既知の語のように扱うことができる。

そこで英語ニュース文の構文解析の前処理として、役職付き人名を複合語として認定する方法を次に述べる。

3．1　役職の認定ルール

英語ニュース文に出現する役職名には,…President (IOC President, U.S. President, etc.)、…Minister (Finance Minister, Foregin Minister, etc)、…Judge (Superior Court Judge, High Court Judge, etc.)等がある。

また、英語では President 1つで 'IOC President'、'U.S. President'などのようにさまざまな役職名を表わすが、日本語では 'IOC会長'、'アメリカ大統領' のよう

Proper Nouns Processing for E-J Machine Translation
Naoto KATOH*, Noriyosi URATANI*, Teruaki AIZAWA*, Sumio NAKASE**
*NHK Science and Technical Research Laboratories
** Catena-Resource Laboratories Inc.

に別の訳語が使われる。したがって、President の訳し分けが必要となる。本ルールではこの訳し分けを President 等の前に付く単語によって行なった。

役職付き人名認定のルールの一部を図２に示す。固有名詞がこの中では複雑に絡みあっている。例えば humanが展開されるときルールの中にも name はあり、company が展開される中にもある。

現在、役職の認定ルールが約１００、役職名が約１３０ほど登録されている。

```
human   －＞ khuman name
khuman －＞ rank
        －＞ FORMER rank
rank    －＞ POST
        －＞ state POST
        －＞ STATE PRESIDENT      (大統領)
        －＞ SOVIET PRESIDENT     (最高幹部会議長)
        －＞ council PRESIDENT    (議長)
        －＞ company PRESIDENT    (社長)
FORMER －＞ former,acting等
POST    －＞ Foreign Minister, etc.
state   －＞ STATE
        －＞ SOVIET
STATE   －＞ U.S., French, etc.   ：国名（ただし、
                        "Soviet"は除く）
SOVIET －＞ Soviet
PRESIDENT－＞ President
council－＞ NCONCIL                ：会議名
        －＞ THE name COUNCIL
NCONCIL－＞ Security Council, etc.
THE     －＞ the
COUNCIL－＞ Council
company－＞ FIRM                   ：会社名
        －＞ name CORP
FIRM    －＞ ATT, etc.
CORP    －＞ Corp., Inc., etc.
```

図２：役職付き人名の認定のルール

３．２　人名の認定

ニュースでは日々変化する全世界の出来事を扱うので、人名、機関名などの固有名詞がすべて辞書に登録されていることは希であり、固有名詞は未知語となることが多い。ところが、未知語であっても３．１で述べたように複合語としてとらえれば、その品詞や意味素性が確定できる。例えば、文中に French President Francois Mitterandと出てきた場合に Francois Mitterand が未知語であっても French Presidentと共に複合語として扱えば、品詞は名詞で'人間'という意味マーカーであると確定できる。したがって、次の段階に行われる構文解析・意味解析でも本質的に既知の単語のように扱うことができる。そこで役職の認定と共に人名の認定も行なった。

人名認定ルールの一部を図３に示す。

この中で未知語があった場合、

１）アルファベットの大文字ではじまり小文字が続く語
　（→ R_NAME）
２）アルファベットの大文字にピリオドが続く語
　（→ R_MNAME）

などを固有名詞の一部とすることにより、未知語となる語を人名とした。

```
name －＞ NAME
      －＞ R_NAME
           (ex. Bush)
      －＞ R_NAME R_NAME
           (ex. George Bush, Overseas Development)
      －＞ R_NAME R_MNAME R_NAME
           (ex. Mikhail S. Gorbachev)
      －＞ R_NAME R_NAME R_NAME
           (ex. Kim Il Sung)
NAME －＞ Geoge Bush, etc.：辞書に登録されている人名
未知語のとき適用されるルール
R_NAME －＞ [A-Z][a-z]+
R_MNAME－＞ [A-Z].
```

図３：人名の認定のルール

４．おわりに

以上、英語ニュースデータベースの簡単な解析結果を検討し、この中で頻出することがわかった役職付き人名を複合語として認定する方法について述べた。

今後は、このデータベースの機能をさらに充実させ、ニュースに頻出する表現（例えば、数字表現・ハイフンを含む単語等）の解析を進めて行きたい。

【参考文献】

[1]相沢他：「衛星放送ワールドニュースの英日機械翻訳」、情報処理学会第40回全国大会