衛星放送ワールドニュースの英日機械翻訳

2F－1 相沢輝昭[1] 江原暉将[2] 浦谷則好[1] 田中英輝[1] 加藤直人[1] 中瀬純夫[3] 有賀憲和[3] 松田健生[3]
([1]ＮＨＫ放送技術研究所 [2]ＡＴＲ自動翻訳電話研究所 [3]カテナリソース研究所)

1　はじめに

ＮＨＫでは、現在、２チャンネルの２４時間衛星放送を実施している。第１チャンネルの中心になっているのは、英、仏、独、伊、露、韓、中の各国語による「ワールドニュース」で、通常は、元のニュースに日本語テロップ（字幕）を重畳して放送している。具体的には、数十人のバイリンガルが、ニュースの聴取、翻訳、要約、原稿の作成までの全てを、限られた時間内で処理している。

しかし、これが衛星放送運用上の大きなネックになっており、その省力化のためＭＴ（機械翻訳）システムが導入された。第１段階として、英語ニュースに対するテロップ作成の実用化を目指す。このため、現在ほぼ毎日５分間、ＭＴシステム作成のテロップを用いた放送が行われている。

2　ワールドニュースと機械翻訳

衛星放送ワールドニュースにおけるＭＴシステムは、バイリンガルによるニュースの聴取、翻訳、要約、テロップ原稿の作成の負担を軽減するために導入されている。まず人間（必ずしもバイリンガルでなくてもよい）がニュースを聴取し、テロップを想定して内容の要約を原文段階で行う。これがＭＴシステムへの入力になる。要約の際、ＭＴシステム向きの前編集をも行う。翻訳結果を見て後編集をしたり、場合によっては原文を修正して再翻訳を行うこともある。

ここで用いているＭＴシステムは、英日機械翻訳システムＳＴＡＲ[1]をベースにしており、基本的にトランスファ方式である。その形態素解析は、単語だけでなく局所的に定形的な単語の並びをも認定するところに特徴を持つ。構文解析では、まず文に対する全ての可能な表層構造を作成し、次に、その中から最適なものを後述するウェイト機構により選択する。現在、辞書は約10万語、文法はCFGタイプの規則で3千弱である。

A Machine Translation System for Foreign News in Satellite Broadcasting by T. Aizawa, N. Uratani (NHK), T. Ehara(ATR), S. Nakase(CLI) et al.

テロップ文（平均11語）の翻訳速度は2MIPSのUNIXコンピュータで２秒以内となっている。

システムは、ニュース文の翻訳用として、さらに以下の特徴を持っている。

3　ニュース用ＭＴシステムの特徴

3．1　ニュース文の特徴

ワールドニュースおよびＡＰ電の約200万語から成る英語ニュース文の特徴を我々は以下のように捉えた。

１）約６万の用語が使われており、しかも分野による分類が難しい。

２）人名、国名、地名などの固有名詞が頻出する。人名は肩書を含んだ複雑な表現が多い。

３）人間を主語に持つ動詞、say, call, report, talk, ask, think, feel等、が多用されている。

４）複雑な数量表現の出現頻度が高い。

５）口語的な表現がよく現れる。

3．2　固有名詞処理

上述のニュース文の特徴１）、２）に対応するために、我々のＭＴシステムではＬＯＣＴ(Local Context Translation)と呼ぶ処理を形態素解析の後段で行っている。その目的は、局所的に固定化している単語列、例えば、

"U.S. President George Bush"

"July 14th, 1789"

"The Metropolitan Museum of Art, New York"

のようなものを認定し翻訳することである。

このLOCTにより、役職付き人名の翻訳、未知の固有名詞の認定、あるいは局所的な情報のみに基づく訳語の選択（例えばpresidentに対する「大統領」「社長」の選択）を、構文解析に大きな負担をかけることなく行っている。LOCTの規則はグローバルな解析規則とは別のCFG規則で記述されている[3]。

3．3　未知語処理

広範なニュース文に対応するため、我々のＭＴシステムでは、大きな辞書に加えて下記の未知語処理機構を備えている。これが、辞書に未登録の語の文法情報を推定し、結果を構文解析部に引き渡す。

1）定形的な単語の並びの認定：　3．2に示したLOCTが、局所的に固定化した単語パターンを認定することによって、その文法情報を推定する。

2）語尾による推定：　英語単語の多くは、その語尾の形態からその文法情報を推定することができる。例えば、-tion(s), -ly, -ing で終る単語は、それぞれ、名詞、副詞、動詞の可能性が高い。

3）特殊ケース：　大文字で始まる語、短い語、数字列については、独自の推定規則を持つ。

3．4　ウェイトを用いた曖昧さの解消

構文解析ははまず入力文に対するＣＦＧタイプの解析を行って、可能な全ての表層構造を１つのＡＮＤ／ＯＲグラフ[2]の形で抽出する。次にその中から最適なものを選択するとともに、それをトランスファに必要な中間構造に変換する。そのための選択手段として用意したのが以下に述べるウェイト機構である。

認定された各単語にはウェイトという評価値が辞書（やLOCT処理）から与えられる。各構文規則にもウェイトが与えられている。例えば、動詞句ＶＰを他動詞ｖｔ２、間接目的語ＮＲ、直接目的語ＮＰに展開する規則には、

:10　VP --> vt2　2: NR　NP

のようなウェイトが与えられている。先頭の10はこの規則の適用に対するウェイトを、2はこの位置の間接目的語の展開に伴うウェイトを示している。ウェイトは、それぞれの単語、句、文の「不自然さ」「内容の煩雑さ」「理解困難性」のようなものを表現する数であり、3．5で述べる意味マーカーの適否によってもその値が変化する。ニュース文の特徴を考慮して、全ての単語と構文規則に対してウェイトを与え、ウェイトが最小となるものを選択している。我々の実験によれば、このウェイト機構の成功率は約７８％である。

3．5　意味マーカーによる訳語選択

訳語の選択手段として、上記のウェイトの外、意味マーカーを使用している[4]。現在のところ、以下の部分でその効果が現われている。

1）ｔｈｅｙの訳し分

ニュース文におけるtheyの出現頻度は高い。しかも、2つの訳語「彼ら」と「それら」の訳し間違いは訳文の品質を著しく低下させる。そこで我々は、以下のような単純な処理で訳し分けを行っている。

3．1に示したように、ニュース文には人間を主語に持つ動詞が多用される。一方、meltのような動詞の主語は、ほとんどの場合意志のない物体である。

そこで、theyの訳語のうち「彼ら」の方には[HIWILL]（高い意志を持つもの）という意味マーカーを付与し、「それら」の方には付与しない。その選択は、動詞の辞書意味記述の主語選択制限に[HIWILL]を指定するか否かによって行える。

2）基本動詞の訳し分け

ニュース文においても多用される基本動詞には実に多様な意味がある。これを訳し分けるには部分的であっても意味マーカーに頼らざるをえない。そのため、ニュース文用のいくつかの意味マーカーを導入した。例えば[CRIMINAL]は、次のような文：

"The police caught the assailant,
who has a history of mental illness."

のcaughtを「逮捕した」と訳すために、その目的語assailantに付けられるものである。

4　結果と考察

衛星放送ワールドニュースの1,393文に対する翻訳結果は次の通りである。

解析に成功したものは898文（全体の64.5%）で、最適な訳文が第１候補にあるもの698文(78%)であった。ウェイト機構による選択はひとまず成功していると言えよう。解析に失敗したものの約30%は原文のスペルないしは文法のミスが原因となっている。口語的表現も解析を難しくしている。

翻訳例

"Mrs. Nishi, with the help of a lawyer, is trying to collect workman's compensation for her husband's death."

「１人の弁護士の助けがあるNishi夫人が、彼女の夫の死のために労働者災害補償を集めることを試みている」

今後は、翻訳精度の向上とともに、ニュース文の特徴４）、５）への対応も進めていく。

参考文献　[1]中瀬「英日機械翻訳システムにおける解析手法について」情処学会NL69-7(1988).

[2]A.Martel et al."Optimizing decision trees," CACM, 21(12) (1978).

[3]加藤ほか「英日機械翻訳における固有名詞処理」本大会(1990).

[4]田中ほか「類語国語辞典を介した意味マーカー付与」本大会(1990).