# 定型パターンを含むニュース文の抽出とその英日機械翻訳

2P－4

加藤　直人　　浦谷　則好

NHK放送技術研究所　　ＡＴＲ自動翻訳電話研究所

## １．はじめに

ニュース用英日機械翻訳システムの精度向上のために、ＡＰ電のニュースをデータベース化し、英語ニュースの分析を進めている。その第一歩として２～１０語の連続して出現する語の頻度をとった。

本稿では、このデータの中である頻度以上出現する連続語を使って、定型パターンを含む文（定型文）を抽出する方法について述べる。このような定型文は従来の機械翻訳のような解析—変換—生成という過程を経て翻訳するよりも、これらの文のみを特別な方法で翻訳する方が精度の向上が期待できる。そこで、機械翻訳の前段階で定型パターンを含む文を翻訳する方法を提案する。

## ２．定型文の抽出

我々は以前、あらかじめ数字や曜日などの同一化処理を行なった後、２～１０語連続の単語列に関する頻度データをとることにより定型パターンを抽出した。さらに定型パターンを使い、１１語連続データ、１２連続データ、・・・を作り定型文を推定した[1]。この方法では文中に定型文を多く含んでいる文であっても、１語でも低頻度の語が途中に含まれていると、定型文として推定できない。

そこで、ある頻度（今回は３回）以上出現した２～１０連続語をデータとし、これらの連続語が１文中に何パーセント含まれているか（含有率＝連続語の語数／総語数）を各文に対して計算した。この中からあるパーセント以上（今回は８０％）の文を定型文として抽出した。本方法では連続語を使うことにより頻度の低い語を含んでいても、その語の前後が定型パターンであれば定型文が抽出できる。また、１単語の高頻度語を使わないことにより基本的な単語（定冠詞、前置詞等）のみが含まれいて、定型パターンを含まない文を排除できる。

本方法で約２年間分のＡＰ電（総文数約１６０万文）から約９.８万文の定型文を抽出した。一部の例を図１に示す。ここで、先頭の数は含有率を表す。

1.000;The U.S. dollar opened at 159.97 yen on the Tokyo foreign exchange market Monday, up from last Friday's close of 157.65 yen.
1.000;The Federal Reserve Board's index measuring the value of the dollar against 10 other currencies weighted on the basis of trade was 97.46 Tuesday, off 0.74 points or 0.74 percent from Monday's 98.20.
0.970;The average price for strict low middling 1 1-16 inch spot cotton declined 99 points to 78.64 cents a pound Wednesday for the seven markets, according to the New York Cotton Exchange.
0.952;The Nikkei Stock Average closed at 25,194.10, down 48.30 points, or 0.19 percent on the Tokyo Stock Exchange Wednesday.
0.852;Philippine peso banknotes Friday at 20.50-21.00 pesos (dealer buying-dealer selling) per U.S. dollar at the close, unchanged from a day earlier.

図１　抽出された定型文の例

## ３．定型文の機械翻訳

２.で抽出した定型文の上位の方には経済ニュースが多かった。経済ニュースの翻訳は非常に難しい。これは、経済独特の英語表現が含まれ複雑な構文を持つ、日本語訳にも経済固有の訳が要求される、等の理由からである。しかし、経済ニュースは数量表現、曜日などが変化するだけという文も多く、一般のニュースに比べると文の種類ははるかに少ない。したがって、経済ニュースに対しては従来のような機械翻訳は必要なく、単純な訳語の置き換えで済む場合が多い。

我々は機械翻訳システムに定型文のみを先に処理する機構を取り入れた。図２に示す。

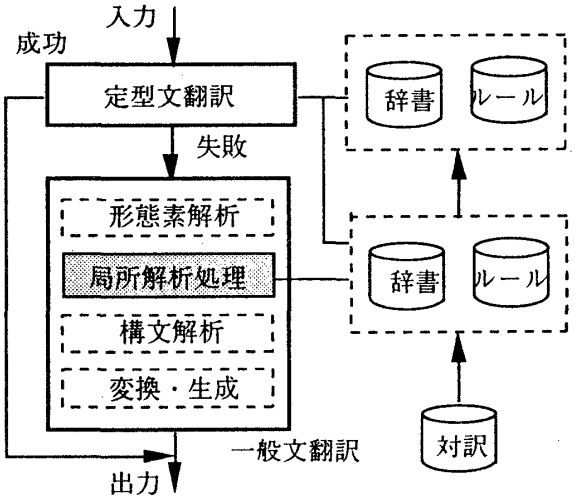


図２　機械翻訳システム

Extraction and machine translation of news sentences with fixed patterns
Naoto KATOH* and Noriyoshi URATANI**
*NHK Science and Technical Research Laboratories
**ATR Interpreting Telephony Research Laboratories

### 3.1 局所解析処理

定型文翻訳に深くかかわる局所解析処理について、簡単に説明する。我々の翻訳システムでは、数量表現、時制表現、固有名詞等は構文解析に先立って、局所的に処理が行なわれる[2]。例えば、

例文1 "In Kuala Lumpur, Malaysian tin closed at 17.76 dollars per kilo, up 5 cents"

という文が入力されると、数量表現は

"17.76 dollars per kilo"　「17.76 ドル／キロ」
"5 cents"　「5 セント」

と局所的に翻訳される。

### 3.2 定型文翻訳

英語の解析はＣＦＧルールで表わした。しかし、英文法に基づくものでなく、単なるパターンマッチングとしている。日本語生成は対応する日本語訳をあらかじめ与え、数量表現等変化する部分のみを置き換える。例えば、例文1の場合には数量表現以外の部分は定型パターンとし、英語の解析は図3のようになる。

```
S --> PAT1 CMA PAT2  UNTEXP CMA UPDW UNTEXP
                                        (1)
PAT1 -->"In Kuala Lumpur"
CMA -->","
PAT2 --> "Malaysian tin closed at"
UPDW --> "up"
UNTEXP --> "17.76 dollars per kilo", "5 cents"
(UNTEXPは局所解析処理による)
```

図3　例文1の英語解析

対応する日本語訳は、

「クアラルンプールでマレーシアのすずは、#6##5#の#4#でひけた」

(#n#は(1)式右辺n番目の記号に対応する日本語訳)

局所解析用辞書に

UPDW -> down　「ダウン」

と登録されていると、例文2も翻訳できる。

例文2 "In Kuala Lumpur, Malaysian tin closed at 18.49 dollars per kilo, down 18 cents"

「クアラルンプールでマレーシアのすずは、18 セント ダウンの18.49 ドル／キロでひけた」

### 3.3 定型文の登録

本手法では定型文の数だけ英語解析用ルールを作る必要があり、日本語訳の中で変化する部分を修正しなければならない。しかし、人手によってこのようなルールを大量に作るのは煩雑である。一方、文法を考慮していないので、このようなルールを自動的に作るのは比較的簡単である。我々は英文に対して対訳の形で日本語訳を与え、局所解析処理を用いて自動的に定型文翻訳用辞書とルールを作成した。アルゴリズムをつぎに示す。

[アルゴリズム]

文頭の語から順に文末まで、単語列に対して

STEP1　局所解析処理が成功したときは、最後に適用されたルールの右辺の非終端記号にその単語列を変更する。その際生成された日本語が、対訳の日本語の中にあれば、変更された記号位置の番号にその日本語を置き換える。局所解析が失敗したならばSTEP2へ。

STEP2　局所解析辞書に登録されているならば、その前終端記号に変更する。その日本語が対訳の日本語のなかにあれば、記号位置の番号に置き換える。辞書に登録されていないならば、STEP3へ。

STEP3　パターンの一部とする。つぎの語に移り、STEP1へ。

例文1では図4のようにルールが自動作成される。

```
1. S --> "In Kuala Lumpur, Malaysian tin closed at 17.76
dollars per kilo, up 5 cents"
「クアラルンプールでマレーシアのすずは、5 セント
アップの17.76 ドル／キロでひけた」
2. S --> PAT1 ", Malaysian tin closed at 17.76 dollars per
kilo, up 5 cents"
3. S --> PAT1 CMA " Malaysian tin closed at 17.76 dollars
per kilo, up 5 cents"
4. S --> PAT1 CMA PAT2 "17.76 dollars per kilo, up 5
cents"
5. S --> PAT1 CMA PAT2 UNTEXP ", up 5 cents"
「クアラルンプールでマレーシアのすずは、5 セント
アップの#4#でひけた」
6. S --> PAT1 CMA PAT2 UNTEXP CMA "up 5 cents"
7. S --> PAT1 CMA PAT2 UNTEXP CMA  UPDW "5
cents"
「クアラルンプールでマレーシアのすずは、5 セント
#5#の#4#でひけた」
8. S --> PAT1 CMA PAT2 UNTEXP CMA  UPDW
UNTEXP
「クアラルンプールでマレーシアのすずは、#6##5#の
#4#でひけた」
```

図4　例文1の英語解析と日本語変換

### 4. おわりに

以上、定型パターンを含む文の抽出とその機械翻訳について述べた。例文1を含む記事の場合、１０文程度の対訳を登録すれば約２年間の文をカバーすることができた。また、定型文翻訳処理では、パターンが登録されているかどうかは早い段階で決まるので、ここでの処理時間は非常に少ない。

今後はスポーツニュース、一般のニュースへの拡張を検討していく。

[参考文献]

１）浦谷他「ＡＰ電経済ニュースからの定型パターンの抽出」情報処理学会第42回全国大会(1991)

２）加藤他「英日機械翻訳における固有名詞処理」情報処理学会第40回全国大会(1990)