ＡＰ電経済ニュースからの定型パターンの抽出

6E−4

浦谷　則好　　加藤　直人　　相沢　輝昭

ＮＨＫ放送技術研究所

1　はじめに

これまでに報告しているように放送用の英日機械翻訳システムの研究を進めている。[1]開発中のシステムは１９８９年８月以降衛星第１放送において日本語テロップの作成に試用されている。翻訳の精度の向上のために，辞書・文法の改良のほかに，定型的なパターンを直接日本語に変換する方式との結合（ハイブリッド方式）も検討している。このために，ＡＰ電の英語ニュースから定型パターンの抽出を行った。

2　ハイブリッド方式

英語ニュース文には，例えば

In Kuala Lumpur, Malaysian tin closed at 28.43 dollars per kilo, down 3 cents.

などの，定型的なパターンがよく出現する．このような文は解析－変換－生成という手順を踏まなくても，「クアラルンプールでマレーシアのすずは１キロ当り，＃Ａドルで引けた。＃Ｂセントの＃Ｃである。」という，日本語パターンを用意しておけば，＃Ａに28.43，＃Ｂに３，＃Ｃに「下げ」を（upなら「上げ」）代入すれば，高品質の日本語訳文を高速に得ることができる。もちろん，Kuala Lumpur, kilo, Malaysian tinも変数として考えることができる。ＡＰ電のデータをもとに，このような定型パターンの利用の検討を進めている。上例のように文全体がパターンとなり得ることもあるが，一般的にはパターンは文の一部として現われると考えられる。我々は，解析－変換－生成を基本にし，文中のパターンは一括して捉え，解析のあいまいさを減らし，高速化と訳文品質の向上を図っていく予定である。これを，ハイブリッド方式と呼ぶことにする。この考え方の一部はすでに，固有名詞や数量表現の処理に利用している。[2]

3　ＡＰ電経済記事に出現する定型パターンの抽出

ＡＰ電１年分（’８９．４～’９０．３：ただし，データの欠落等のため実質約３００日分）の経済ニュースからパターンを抽出するために，まず単語の連続出現頻度を調べた。長いパターンを抽出するにはできる限り長い単語列の頻度を取った方が良いが，記憶容量と計算時間を考慮して，１０語連続を調べることにした。テキストから１語づつずらして全ての１０語連続の頻度を求めた。こうしておけば，９語連続以下の頻度情報も求めることができるし，これくらいの長さがあれば，１１語以上の頻度も後述する方法で推定することができる。ただし，文末を越えての語の連続は認めないこととし，コンマや括弧等も１語と見なすことにした。さらに，重複記事と判別できるものは統計対象から除外した。結果として得られた頻度表を表１に示す。一番出現している"futures trading on ..."は612回（全体（145万）の0.04%）で，１日に平均２回程度出現していることになる。

表１．　１０語連続の出現頻度（経済記事）　　（上位１０位）

futures trading on the New York Coffee , Sugar and	612
trading on the New York Coffee , Sugar and Cocoa	611
on the New York Coffee , Sugar and Cocoa Exchange	608
a bid price of ¥ dollars a troy ounce ,	349
(dollars per metric ton) cash ¥-¥ (¥-¥	345
at a bid price of ¥ dollars a troy ounce	298
in London at a bid price of ¥ dollars a	264
London at a bid price of ¥ dollars a troy	264
, with ¥ up , ¥ down and ¥ unchanged	258
and dealer buying-dealer selling U.S. dollars at ¥-¥ Hong Kong	247

［¥：　数字列］

3．１　１０語連続頻度表からの長短頻度表の作成

表１からも分かるように，１０語連続は実はもっと長いパターンの部分である可能性がある。（上位３個は１２語以上のパターンの一部と考えられる。）また，１０語ではパターンをなさないが，９語以下でパターンとなっているものも考えられる。パターン表現としてはできるだけ長いものを抽出したい。また，曜日表現など変数として扱うべきものを個別に扱っていては，パターンの数が増えるだけで有効なパターンが現われてこない。そこで，同一と見なした方が良いと思われる単語群を，曜日を全てWednesday，月は通常表現はAug.，フルスペルのものはNovember，記号的表現はAprにするなどの同一化を行うことにした。短いパターンの抽出のために，１０語連続頻度表から後ろの１語を順次除いてゆき，９語・８語・７語・６語の頻度表

Extraction of fixed patterns from AP economy news
Noriyoshi URATANI, Naoto KATOH and Teruaki AIZAWA
NHK Science and Techinical Research Laboratories

を作成した。また，長いパターンを求めるために，１１語，１２語，・・・の頻度表を以下のように作成した。

＜ｎ＋１語頻度表の作成＞

ｎ語のある単語列α(例えばabcdefghij)の後ｎ－１語と別の単語列β(bcdefghijk)の前ｎ－１語が一致したときは，一致した部分(bcdefghij)を含むｎ＋１語の単語列γ(abcdefghijk)を生成する。この単語列の頻度はα，βの内小さい方の値とする。ただし，α，βの頻度の差が大きいとき（片側が２倍以上）にはｎ＋１語単語列を生成しない。

３．２　定型パターンの自動抽出

あるｎ語の単語列の頻度がｍであれば，前後どちらかの語を除いたｎ－１語の単語列の頻度は必ずｍ以上となるから，単純に頻度が大きいからといってそれをパターンと考えるわけにはいかない。大量のデータを扱うときには，真のパターンとなり得るものを何らかの基準をもとに自動的に抽出したい。これには，仕事量の基準[3)]を採用することにした。すなわち，単語列αの総仕事量を

$$W(\alpha) = |\alpha| \times n(\alpha)$$

（ここで$|\alpha|$　は単語列の長さ（語数）

$n(\alpha)$は単語列の出現頻度　）

で定義する。（ただし，文献３と異なって$|\alpha|$を用いて，仕事量の差でなく，仕事量そのものを用いている。）そして，単語列αがあるとき，この先頭の単語を除いた単語列βとし，最後尾の単語を除いた単語列γとしたとき，

$W(\alpha) > W(\beta)$　かつ　$W(\alpha) > W(\gamma)$

のとき，αを有効なパターンとして認定する。

ただし，単語列αを部分列として含む１だけ長い単語列のうち最大頻度を持つものを単語列δとしたとき，αの頻度を$n(\alpha) - n(\delta)$で更新し，これが２以下になるものは除外する。つまり，αはδで表現されないものの代表と考えるのである。

３．１で得られたデータをもとに，上記の基準で有効パターンの抽出を４４語連続まで行った。１１語連続以上で頻度200（0.014%）以上のものを表２に示す。１０語連続以下の上位３位までを表３に示す。

表２．　１１語以上の出現パターン（経済記事）　（頻度２００以上）

１２語	
futures trading on the New York Coffee , Sugar and Cocoa Exchange	608
to close at ¥ U.S. dollars , compared to Wednesday 's ¥	240
１３語	
's foreign currency prices (dearler buying-dealer selling) quoted in U.S. dollar	211
the U.S. dollar closed at ¥ yen on the Tokyo Foreign Exchange Market	208
In Kuala Lumpur , Malaysian tin closed at ¥ dollars per kilo ,	208

表３．　１０語以下の出現パターン（経済記事）　（上位３位）

１０語	
the New York Coffee , Sugar and Cocoa Exchange Wednesday	609
(dollars per metric ton) cash ¥-¥ (¥-¥	345
, to close at ¥ U.S. dollars , compared to	244
９語	
European Currency Unit , a basket of European currencies	205
Futures trading on the Chicago Board of Trade Wednesday	193
of ¥ dollars a troy ounce , up from	184
８語	
--¥ West German marks , up from ¥	358
¥ futures on the N.Y. Cotton Exchange Wednesday	205
¥ dollars a troy ounce , up from	185
７語	
--¥ French frans , up from ¥	381
--¥ Swiss frans , up from ¥	376
--¥ Dutch guilders , up from ¥	373

４　おわりに

表２，表３を見ると表１ではノイズ中に埋もれていたパターンの候補が顕現していることが分かる。しかし，単語の同一化処理の不足のために，本来はもっと長いと推測されるパターンの一部が現われていると思われるものもある。ところで，ｎ語の連続頻度表からｎ－１語が一致するものを取ってきて，異なっている語を調べれば，同一化すべき単語（変数）の候補も得られると考えられる。この結果は別の機会に報告したいと考えている。

ここでは，経済記事についての結果を述べたが，一般記事についても１０語連続頻度を取ってみた。結果は一番頻度の高いものでも116回（全体（1600万）の0.0007%）であった。頻度50以上のものは８個で，一般記事にはそんなに長いパターンは現われないことが判明した。

今後は，得られたパターンの候補を精察し，各パターン処理方法を決めていく予定である。必要とあれば，節・句・品詞などの解析単位の見直しも行う予定である。さらに，スポーツ記事や一般記事（の短いパターン）についても，解析を進めていく予定である。

【　参考文献　】

１）浦谷ほか「英語ニュースの機械翻訳」情報処理研究会ＮＬ７８－１８(1990)

２）加藤ほか「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２Ｆ－２(1990)

３）北ほか「テキスト・データベースからの慣用表現の自動抽出」ATR Techinical Report TR-I-0027