日本語ニュース文の慣用パターンの分析

1H－5　浦谷 則好　畑田 のぶ子
NHK 放送技術研究所

1．はじめに

機械翻訳等の自然言語処理にとって慣用表現や定型表現が重要な役割を果たすことは、疑うべきもない。例えば機械翻訳では定型表現を翻訳ユニットにとれば精度の良い翻訳が期待できる。そこで、コーパスからコンピュータを用いて機械的に慣用表現を抽出することは各所で研究されている[1)~6)]。機械的に慣用表現を抽出するためには何らかの基準が必要となる。我々は３つの基準を設定して表現パターン（慣用表現や定型表現）を抽出する実験を実施し、前回の全国大会で報告した[6)]。前回の結果を用いて日本語ニュース文の定型パターンの分析を行なったので、それについて報告する。

2．ニュース文の特徴

ニュース記事の構成は通常、先に事実記述が来て、後に補完情報が続くという具合になっている。試みにＮＨＫのニュース記事をランダムに１００記事分選び調べてみたところ文の数の分布は図１のようになった。１記事は大体４～７文で構成されていて、第１文には全体の要約が記述される。第２文には第１文を補う形で事実が記載され、第３文以降には大体、理解の助成のための情報（情報源、解説、補足）が述べられている。

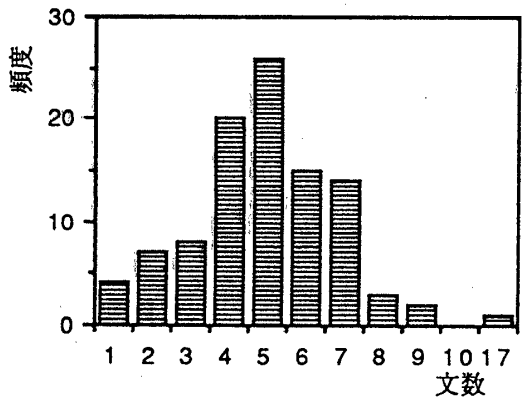

図１．ニュース記事の文の数

前回の実験で、機械翻訳を前提にして定型パターンを抽出するにはエントロピー基準を用いるのがよいことが判明している。そこで、エントロピー基準で選び出した上位約11,000個のパターンから、文をまたぐもの、１字のものを除いた約8,600個について記事中に出現する割合を求めた。文の位置（文番号）に対する文の長さの平均とパターンの被覆率（平均）を図２に示す。これを見ると、第

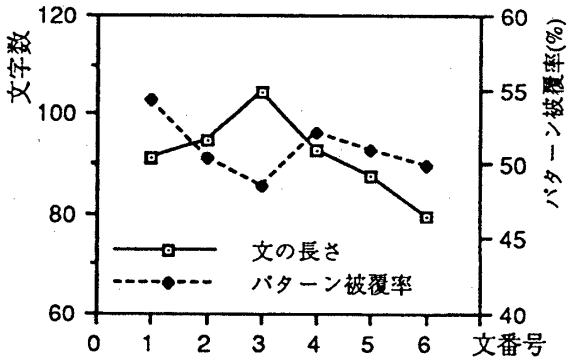

図２．　ニュース文の長さとパターン被覆率

１文はほぼ中間の長さ（９１文字）を持つが、被覆率は高い（５４％）ことがわかる。第３文は文が長く（１０４文字）、被覆率は低い（４８％）。より、詳しく調べるため、パターン被覆率の高

表１　被覆率の違いによる文の特徴量

	文番号	件数	文字数
上位１００文	1	26	102.0
	2	17	82.5
	3	10	76.9
	4	20	80.2
	5	13	81.8
	6－	14	59.0
下位１００文	1	20	68.5
	2	18	93.2
	3	19	89.1
	4	16	89.7
	5	14	77.4
	6－	13	89.4

An Analysis of Fixed Patterns in Japanese News Articles

Noriyoshi URATANI and Nobuko HATADA
NHK Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo 157, Japan
e-mail:{uratani,hatada}@strl.nhk.or.jp

い（６９％以上）１００文と低い（３３％以下）１００文を選び出したところ表１のような結果がえられた。これを見ると、第１文で被覆率が高いのは、文が短いからではないことがわかる。つまり、第１文（＝事実記述）は定型的で、第３文（=理解助成部分）は非定型的であると推論できる。

３．定型パターンの特徴

前述したパターンの文中で占める役割を調べるために、パターンを形態素解析辞書に登録して、形態素解析を実施してみることにした。抽出されたパターンのうち、すでに辞書にあるものや品詞付けの不可能なものを除いて、1170個の単語の登録を行なった。この際、従来には存在しなかった２つの品詞（格動詞と特動詞）を新設した。格動詞とは格助詞と密接に結合している動詞、例えば「との見方を示しました」を１つにまとめたもので、特動詞は「疑いが持たれています」などのような動詞的な文末パターンを扱うもので、「違反の」や「頼んだ」などのような修飾要素が直前に付きうるので本動詞とは区別するために設けたものである。登録したパターンの概要を表２に示す。名詞類の多くは基本語辞書(46,000語)から漏れていたものの補完であるが、一部は名詞句を１語として扱うようにしたものもある。特徴的なものは前述した格動詞や特動詞のほか、格助詞、連体助詞、助動詞である。基本語辞書でも格助詞は66個、連体助詞は16個と通常より多い目に登録されているが、それにニュース特有の表現パターンを追加した。助動詞は通常の日本語解析では形式名詞＋動詞等になる形のもの（例えば「ことが確認されました」、「ものとみられています」など）でも、翻訳を前提にしたときには助動詞的に扱われるものを広く採用した。従来の形態素解析と上述したパターンを加えた形態素解析との相違を以下に例示する。

《従来》　八王子／を／中心（普通名詞）／に／およそ／十／万／人／に／影響／する／もの（普通名詞）／と／見（本動詞）／られ／ま／す／。

《本方式》八王子／を中心に（格助詞）／およそ／十／万／人／に／影響／する／ものと見られま（助動詞）／す／。

これを見ると、定型パターンを利用した方が効率的であり、翻訳ユニットとしても適切なものが得られることが予想される。

表２　パターンを基に追加した形態素

品詞	数	例
サ変名詞	33	最高値を更新
引用助詞	2	かどうかについて
格助詞	47	が中心となって
格動詞	30	という考えを明らかにしました
係助詞	3	としては
形容詞	4	やむを得ない
固有名詞	99	ＥＣ・ヨーロッパ共同体
終助詞	1	のではないか
助動詞	115	ている模様です
人名	21	キム・ジョンイル
数詞	3	シックス
接続詞	33	これまでの調べによりますと
接続助詞	47	としながらも
接尾辞	3	年ぶり
特動詞	125	ことを示唆しました
普通名詞	221	これまでの最高値
副詞	31	小幅ながら
補助動詞	9	してもらう
本動詞	306	検討することにしています
連体詞	8	きょう発表した
連体助詞	29	を含めた

４．おわりに

抽出されたパターンを用いて日英機械翻訳のための翻訳ユニットについて実験結果をまじえ検討した。今後、さらに定型パターンの文中の役目について考察を加えるとともに、文（あるいは節）のパターンについて分析を進めていく予定である。

参考文献

1）Church, K.W. and Hanks, P.:Word Association Norms, Mutual Information, and Lexicography, ACL-89 (1989)

2）浦谷則好ほか：ＡＰ電経済ニュースからの定型パターンの抽出，情処４２全大６Ｅ－４，(1991)

3）北　研二ほか：仕事量基準を用いたコーパスからの定型表現の自動抽出，情処論Vol.34, No.9, (1993)

4）長尾眞，森信介：大規模日本語テキストのｎグラム統計の作り方と語句の自動抽出，情処研資NL96-1, (1993)

5）新納浩幸ほか：コーパスからの関係表現の自動抽出，情処論Vol.35, No.11, (1994)

6）浦谷則好：ニュース原稿データベースからの表現パターンの抽出，情処５０全大１Ｒ－８，(1995)