1R－8

ニュース原稿データベースからの表現パターンの抽出

浦谷 則好
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。しかし、慣用表現を人手で収集するのは容易ではない。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が考えられている[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。過去の研究では相互情報量や仕事量などの基準が使われているが、どのような基準が慣用表現を抽出するのに適しているかは明らかでない。

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た。ここでは、３つの基準と抽出結果と傾向について報告する。

2．表現パターンの自動抽出

ある文字列（長さｎ）が１つの表現パターンをなしているか否かは、直感的に「その文字列の先頭か最後尾の１字を取ったパターンの頻度が元のパターンからあまり増加しない」、および「その文字列の前後に１字を加えたパターンの頻度は（最大のものでも）大きく減少する」ことを確認することで決定することができる。しかし、どの程度の変化をもって表現パターンの大きさを決定すれば良いかは明らかでない。Churchらは相互情報量[1)]を、北らは仕事量[4)]を、長尾らは直前、直後にくる文字種の数[5)]を、新納は条件付き確率の変移[6)]を抽出の基準に用いている。

われわれは以下に示す３つの基準を設定して、抽出実験を行なった。

●仕事量基準

$n \times f_\alpha$

●平均情報量基準

$log\ (N / f_\alpha) / I_n$

●エントロピー基準

$log\ (N / f_\alpha) / H_n$

ここで、

α：長さｎの文字列（例えば defgh）

f_α：文字列αの頻度

A_n：長さｎの文字列全体の集合

N：A_nに含まれる文字列の頻度の総和

f_n：A_nに含まれる文字列の頻度の平均

I_n：A_n中で平均頻度を持つ文字列の情報量

$= -log\ (f_n / N)$

H_n：A_nのエントロピー（$= -\Sigma p\ log\ p$）

である。「仕事量基準」は北らのものと酷似しているが、ｎ－１でなくｎを用いているところが異なっている。（北らの基準では、１字長のパターンと２字長のパターンの優位を計ることができない。また、重要度を定義するにはこの方が自然であると考える。）

これらの基準を用いて、各々の文字列をそれより１だけ長いもの、および１だけ短いものと比較し、優位と判断された方を残すことによって表現パターンを抽出する。例えば、

β：αを部分列として含む長さｎ＋１の文字列（例えば defghi あるいは cdefgh）

γ：αの先頭の１字を除いた文字列（例えばefgh）

δ：αの最後の１字を除いた文字列（例えばdefg）

とした時、「仕事量基準」では

$\max_\beta (f_\beta) \times (n+1) < f_\alpha \times n$

で、かつ

$\max (f_\gamma, f_\delta) \times (n-1) < f_\alpha \times n$

のときにαを表現パターンと認定するわけである。（「平均情報量基準」や「エントロピー基準」でも同様であるが、不等号の向きは反対になる。）

An Automatic Extraction of Expressions from News Manuscript Database

Noriyoshi URATANI
NHK Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo 157, Japan

3．実験

ＮＨＫのニュース原稿データベースを対象にして２．で述べた基準を用いて表現パターンの抽出を行なった。用いたデータは１９９２年７月から１９９３年７月分の約１年分のニュース文（本文のみで１１１０万文字）である。文字列の単純頻度は長尾の方法[5]で求めた。表現パターンの抽出結果を図１～図３に示す。“S¦”は文頭を表す記号で１字として扱っている。これを見ると仕事量基準では短いパターンが抽出される傾向があり、エントロピー基準では比較的長いパターンが抽出される。平均情報量基準は２つの中間に位置しているように思われる。

4．おわりに

抽出されたパターンは３つの基準それぞれに特徴を有している。我々の目的は日英機械翻訳のための表現パターン（つまりは翻訳ユニット）を探ることにある。実験結果から、こうした目的にはエントロピー基準が合致しているように思われる。今後、さらに実験と検討を加えて、日英機械翻訳システムの改善に利用していく予定である。

表１　仕事量基準で抽出されたパターン

仕事量	頻　度	表現パターン
484380	484380	の
320183	320183	、
299345	59869	ました。S¦
263361	263361	に
240718	120359	。S¦
218431	218431	と
214180	107090	した
213325	213325	を
197914	197914	で
184302	30717	ています。S¦
182450	182450	が
179476	89738	てい
166410	55470	す。S¦
162040	81020	して
123046	61523	は、
93098	46549	こと
92178	46089	って
86902	43451	する
84002	42001	から
82491	27497	ている

表２　平均情報量基準で抽出されたパターン

頻　度	表現パターン
59869	ました。S¦
30717	ています。S¦
55470	す。S¦
484380	の
89738	てい
81020	して
61523	は、
18336	について
24091	。S¦こ
21987	ること
263361	に
43451	する
42001	から
18051	きょう
17921	という
33935	など
6798	によりますと
14024	ことに
213325	を
8713	アメリカ

表３　エントロピー基準で抽出されたパターン

頻　度	表現パターン
59869	ました。S¦
30717	ています。S¦
18336	について
17665	です。S¦
4403	ということです。S¦
6798	によりますと
6017	ということで
3045	ことにしています。S¦
27497	ている
2894	ことになりました。S¦
3781	ました。S¦また
3172	を示しました。S¦
3585	たものです。S¦
2543	と話しています。S¦
484380	の
2428	」と述べました。S¦
5827	す。S¦この
21987	ること
21587	として
2982	」と話していま

参考文献

1）Church, K.W. and Hanks, P.:Word Association Norms, Mutal 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）加藤直人，相沢輝昭：外電ニュースの定型文抽出とその英日機械翻訳，情処研資NL93-2, (1993)

4）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)

5）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)

6）新納浩幸：文字列と後続文字との接合割合の変化を利用した定型的文末表現の自動抽出，情処研資NL104-6, (1994)