日本語ニュース文の表現パターンの分析

浦谷則好　　畑田のぶ子
NHK 放送技術研究所
{uratani,hatada}@strl.nhk.or.jp

1．はじめに

自然言語処理にとって慣用表現や定型表現の処理が重要である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。しかし、こうした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は各所で研究されている[1)~5)]。機械的に定型表現を抽出するためには何らかの基準が必要となる。我々はエントロピー基準を用いることで機械翻訳のための表現パターン（翻訳ユニット）を精度良く抽出できることを実験によって確認している[6)]。ＮＨＫニュース原稿から抽出した上位8,600個のパターンがニュース文全体の約50%を被覆することも確認している[7)]。これまでの結果を用いて日本語ニュース文の表現パターン連鎖型の共起と離散型の共起を調査したので、それについて報告する。

2．対象パターンの選定

１９９２年7月から１９９３年7月までの約1年分（実質約３００日分；１１１０万文字；１２万文）のＮＨＫのニュース原稿からエントロピー基準を用いて表現パターンを抽出した。その中からニュース文の特徴を示しているとは思えないパターン（「ました。」「です。」「ている」「の」など）を除いて、上位から名詞類は頻度２００以上のもの、それ以外は頻度１００以上のパターンを人手により全部で６０７個選定した。選定したパターンの例を表1に示す。これを対象に連鎖型の共起と離散型の共起を調査した。まず、長尾・森の方法[4)]（ただし文番号を保持するファイルを加えた）を用いて文字列とその出現位置、文番号を抽出し、パターン毎にインデックスファイルを作成した。次に、その情報をマージして文毎にどういうパターンが含まれているかを示すファイル（「文内パターンファイル」と呼ぶ；図1）を作成する方法を採った。池原らの方法[8)]を採用しなかった理由は汎用ソートファイルを作成するより、文内パターンファイルを作るほうが簡便だと考え

表1　選定した表現パターン

分　類	具体例	数
格助詞的表現	によりますと に対して	６５
名詞的表現	政治改革 エリツィン大統領 警察本部	１２７
動詞的表現	ことにしています と話しています 明らかにしました 強調しました	２７７
副詞的表現	さらに その後	４８
そ の 他	のではないか べきだという ているもので	９０

文番号j　開始位置j_1　パターンj_1
　　開始位置j_2　パターンj_2
　　：　　：
　　開始位置j_n　パターンj_n

＜例＞
2　９０　によりますと
　１６５　と呼ばれる
　１９７　というもので
5　３３９　これまで
　３７０　となってい
6　５６２　選挙制度改革
　５９８　を受けて
　６１０　提案し

図1　文内パターンファイルの構成

たからである。選定したパターンの中には「によりますと」と「それによりますと」のように一方が他に完全に含まれてしまうものも存在している。この場合、もし長い方のパターンが出現した場合は短い方はカウントしないようにし、重複して数えることを避けた。全文のうち選定したパターンを1つも含まないものは16．6％しかなかった。また、パターンを1つしか含まないものは27．6％であった。残り55．8％は2つ以上のパターンを含んでいることになる。

3．連鎖型の共起パターンの抽出

2．で選定したパターン相互の位置関係には図2に示す2つのケースがありうる。すなわち、

（a）パターンαとパターンβが部分的に同一文字列を共有する場合　と

（b）パターンαとパターンβが文字列を共有しない場合

である。このうち、（a）を連鎖型の共起パターン、（b）を離散型の共起パターンと呼ぶことにする。2つのパターンが連鎖型で共起しているか、離散型で共起しているかは文内パターンファイルのパターン相互の出現位置とその長さを調べることで簡単に調べることができる。

調査したデータ中、連鎖型の共起パターンを形成する場合、最大5つのパターンが連鎖することが分かった。それは1例だけで「のではないかという見通しを示しました」というパターンで、頻度は8であった。元になったパターンは「のではないか」、「ではないかと」、「という見通しを」、「見通しを示し」、「を示しました」の5つである。

パターン2つで連鎖型の共起パターンとなるのは820例（頻度総計15,719）、3つの場合は270例（頻度総計1,768）、4つの場合は35例（頻度総計81）であった。頻度の多い順に連鎖パターンを示すと表2のようになる。線で区切って上から2，3，4つのパターンから成る場合を分

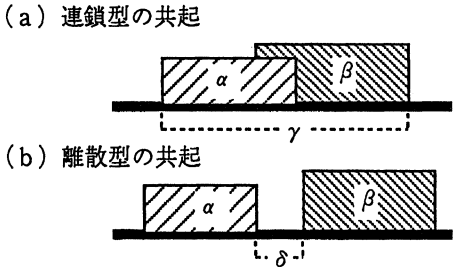

図2　パターンの共起型の分類

表2　連鎖型の共起パターン（上位）

パターン	頻度
これに対して	715
のではないかと	597
ことを決めました	450
ことを明らかにしました	434
という考えを示しました	428
べきだという考えを	423
のに対して	359
の記者会見で、	283
ものと見られています	249
ていることから、	231
ことになりそうです	207
警察の調べによりますと	198
たいという考えを示しました	184
ことが明らかになり	178
調べることにしています	172
疑いで逮捕しました	164
手当てを受けて	158
供述しているということで	150
これまでの調べによりますと	148
を進めることにしています	142
姿勢を示しました	139
警視庁の調べによりますと	137
行方不明になってい	136
べきだという考えを示しました	185
べきだという考えを強調しました	110
午前の記者会見で、	87
という見通しを示しました	69
ているのが見つかった事件で	59
殺人の疑いで逮捕しました	52
になるという見通しを示しました	20

けて示してある。これを見るとエントロピー基準では分離されてしまっても、頻度が多いものは意味のある単位で連鎖型の共起パターンとして抽出されることがわかる。

4．離散型の共起パターンの抽出

3．で抽出した連鎖型の共起パターン（図2（a）のγに相当：１１２６種）も選定パターンと見なし、離散型の共起パターン（図2（b））の抽出実験を行なった。文内パターンファイルから離散型の共起パターンを距離（パターンの間の文字数；図2（b）のδに相当）まで含めて抽出し、頻度の大きいもの順にソートした。結果を表3に示す。

表3で距離1の場合に入る文字はほとんどの場合、記号類である。第6位のパターンを見てみると、距離0で頻度は１００を超えている。それならば「なぜ最初に選定したパターンに含まれなかったのか？」という疑問が生じる。この理由は、「の調べによりますと」の前にくるパターンは「警察本部」ばかりでなく、連鎖型共起パターンに出てくるように「警察」や「警視庁」などがあるので、エントロピー基準では「警察本部／の調べによりますと」と分離されてしまったからである。頻度の高い距離0の離散型共起パターンや連鎖型共起パターンは、エントロピー基準で分離されてしまうパターンを再結合する機能を持っていると言える。

第4位の間に入る文字列は「（は）一億九千万」などの数値である。このことから、こうしたパターンは単純に穴埋め型のパターン翻訳に利用できることがわかる。

次に距離を無視して離散型の共起パターンを抽出した結果を表4に示す。これを見るとニュースに現われやすい呼応がうまく抽出されていることがわかる。翻訳パターンとしては、例えば「Aによりますと Bということです。」に対して"According to A', B'"、あるいは"A' reported that B'"のような対応を考えていけばよいことが

表3　距離別の離散型共起パターン
（上位１０位）

前方パターン／後方パターン	距離	頻度
セルビア人／武装勢力	0	１７８
ＰＫＯ／国連の平和維持活動	1	１５０
取り調べ／に対して	0	１２４
一日の出来高／株でした	6	１１６
主要銘柄の平均株価／午前の終値は	1	１１３
警察本部／の調べによりますと	0	１１１
に対する／警戒感	0	１０６
円相場／午前の終値は	1	１０２
ていきたい／と話しています	1	８９
ていきたい／と述べました	1	８７

表4　距離を無視した離散型共起パターン
（上位２０位）

前方パターン／後方パターン	頻　度
市場関係者は／と話しています	５５０
この中で／と述べました	５３３
によりますと／ということで	４６４
の調べによりますと／ということで	２９９
これについて／と話しています	２８０
気象庁／によりますと	２３０
予算委員会／証人喚問	２１９
セルビア人／武装勢力	１９７
東京佐川急便／証人喚問	１８１
この中で／としては	１７７
に対して／するよう	１７４
この中で／を示しました	１７１
金丸前自民党副総裁／脱税事件	１６８
の話によりますと／ということで	１６５
野党側／証人喚問	１６３
ＰＫＯ／国連の平和維持活動	１５９
それによりますと／としてい	１５９
この中で／に対する	１５８
ボスニア・ヘルツェゴビナ／セルビア人	１５７
ラナリット／プノンペン政権	１５６

予想される。名詞同士の共起（例えば「野党側」と「証人喚問」）の情報は、ニュース分野の特定や訳語の選択に利用できると思われる。

5．文パターンの抽出

3．や4．から、連鎖型や離散型のパターン共起の抽出がニュースを対象とした自然言語処理に有効であることがわかった。しかし、パターン間の共起を調べていくと選定したパターンの中には他のパターンとほとんど共起しないものがあることも判明した。例えば、「逮捕されたのは」は単独では５２２回も出現するが、他のパターンとの共起では「を含む」の６回が最高である。

そこで、「逮捕されたのは」を含むニュース文を人手で調べて、本当に文としてはパターン化されていないかを調べた。すると、「（再）逮捕されたのは～です。」というパターンと言えるものが４７７例もあることが分かった。これが、パターンとして捉えられなかったわけは、2．で「です。」を選定パターンから排除してしまったためである。また、「（再）逮捕されたのは～です。」というパターンはより詳細に見ると実は"～"に関しても極めてパターン化されていることがわかった。人手によって分類した結果を表５に示す。

「{～た¦～る}のはＸ初めてです。」というのもニュースでよく現われる表現である。対象データ中には１４５回出現している。ここでＸに入りうるものを列挙してみると、「これが」（69）、「ε（何も無し）」（39）、「今回が」（31）、「全国でも」（6）となっていた。文パターンの抽出にはこのＸのような変数の選定が重要である。

6．おわりに

エントロピー基準を用いて抽出したパターンからニュース文中の連鎖型および離散型のパターンの抽出を行った。連鎖型の共起パターンはエントロピー基準により抽出したパターンの不備を補うのに有効なことがわかった。離散型の共起パターンは文としてのパターンの把握やニュース分野の推定などに有効であることが判明した。今後、さらに表現パターンの文中の役目について考察を加えるとともに、5．でふれたように変数の導入によって文（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。

表５　「（再）逮捕されたのは～です。」の表現パターン　（～部分のみ示す）

<住所>{の¦に住む}<職業>(の<肩書き>)<人名>容疑者（～歳）です。	８５
(<住所>生まれの)住所不定・無職の<人名>容疑者（～歳）です。	１４
<住所>の{～会系¦～組系}暴力団(～組)<肩書き><人名>容疑者（～歳）です。	１０
<住所>{の¦にある}(～会社)<企業名><肩書き><人名>容疑者（～歳）です。	１０
<住所>の<職業>、<人名>容疑者（～歳）と、<住所>の<職業>、<人名>容疑者（～歳）の二人です。	９
<住所>{の¦にある}<職種>会社の<肩書き><人名>容疑者（～歳）ら(あわせて)？人です。	３

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，井佐原均：疑似Ｎグラムを用いた助詞的定型表現の自動抽出，情処論Vol.36, No.1, (1995)

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)

7）浦谷則好，畑田のぶ子：日本語ニュース文の慣用パターンの分析，情処５１全大１Ｈ－５，(1995)

8）池原悟ほか：大規模日本語コーパスからの連鎖型および離散型の共起表現の自動抽出法，情処論Vol.36, No.11, (1995)