放送ニュース文における頻出パターンの調査

西脇 正通　金城 由美子　柏岡 秀紀　田中 英輝
ATR音声言語通信研究所

1. はじめに

著者らは放送ニュースや講演などの独話を対象とした同時通訳手法を研究している。放送ニュースにはさまざまな話題が出現するため、表現が多様となり精度良く自動翻訳することは難しい。このため翻訳対象のドメインを自動翻訳に適したドメインに限定したい。放送ニュースのドメインを限定する一つの方法に、あらかじめ与えられている分野分類を使って対象を絞ることが考えられる。しかし、この選択によって表現の多様性が減少するとは限らない。

そこで本稿では、ニュースに出現する表現を節単位で抽出して、その多様性や出現の偏り方について調査した。また、経済分野に限ったときに表現の多様性がどのように変化するかを調査したので報告する。

2. コーパスからの表現パターンの抽出

2.1. ニュース・コーパス

研究に際しては、放送ニュースの原稿を蓄積したニュース・コーパス[i][1]を使用し、放送ニュースの特徴的なパターンを抽出した。

コーパスは、原稿を作成した部局をもとに各分野に分類されている。この分野分類を「ジャンル」と呼ぶ。

1995年4月から1998年3月までのコーパスに含まれる原稿のジャンルとそれぞれの原稿数を表1に示す。

表 1　コーパスに含まれる原稿のジャンル

ジャンル	経済	国際	政治
原稿数	16,347	20,476	24,269
ジャンル	社会	スポーツ	その他
原稿数	30,241	16,946	55,810
合計原稿数	164,089		

2.2. 日本語解析

本稿では放送ニュースの表現を調査するために、次節以降に示す表現パターンを抽出した。パターンの抽出のために、放送ニュース文に日本語解析[ii]を行ない、文節間係り受け構造をもつ構文データーを作成した。

表2に対象としたコーパスの範囲と構文データーの概要を示す。

表 2　パターン抽出対象データの概要

範囲	1995年4月～1998年3月（1995年12月を除く35ヶ月）[iii]
原稿数	144,196[iv]
文数	791,051[v]　（5.5文/原稿）
文節数	12,999,253　（16文節/文）
形態素数	39,588,349　（3.0形態素/文節）

2.3. パターンの認定

文節間係り受け構造から抽出するパターンは、1つの述部を持つ節単位とした。ただし、連体節の場合は修飾先の体言もパターンに含み、その場合は係り受け関係にあるパターン間に体言の重なりを認めた。

また、途中で分割すると言いまわしとして不自然なものは、複数のパターンに分割しなかった。表3に分割が不適当と考えた例を示す。分割しないと判断するために下線部分の単語を用いている。

表 3　複数のパターンに分割しなかった例

話し合われるとの見通しを示しました
見つめていると思っていると述べるにとどまりました
編成する必要はないと述べ

2.4. パターンの汎化

形態素解析で品詞が数字になったものは「NUM」に、固有名詞になったものは「PNOUN」に置き換え、パターンを汎化した。数字または固有名詞が隣接する場合や、固有名詞が「・」を挟んで隣接する場合はまとめてひとつの「PNOUN」や「NUM」に置き換えた。

2.5. 句読点、括弧、記号の省略

パターンの同一性を確認する時、「首相は、述べました」と「首相は述べました」は同一のパターンとしたいので、抽出されたパターンから読点を削除した。また、1文内における出現位置にかかわらずパターンを評価したいので、句点も同様に削除した。その他にも、一部の記号は削除している（表4）。

表 4　削除した文字

。 、 「 」 ▽ ▼ △ ▲ ○ ！ ？ ， ． 『 』 ”

i 研究利用のためNHKから提供をうけたもの
ii 日本語解析には、研究利用のためNHKから提供をうけた日本語解析器を使用した。
iii 1995年12月はデーターが欠損しているため対象となっていない。
iv 本文が無い19,893の原稿をあらかじめ除外している。
v 800文字以上の67文は調査対象文として認められないものがほとんどであったので、あらかじめ除外している。

3. パターンの偏りの評価

3.1. クロスバリデーション

月単位でのパターンの偏りを探るために、35回のクロスバリデーションを行ないカバレッジを比較した。35回のクロスバリデーションでは、35ヵ月分のコーパスを月単位で分割し、34ヶ月分の標本データーとそれに含まれない1ヶ月分の評価データーの組を使用した。同様に、コーパスを時系列順に7ヶ月単位で5分割した5回のクロスバリデーションでリコール、プレシジョン、カバレッジを比較した。

3.2. パターンの頻度とリコール、プレシジョン、カバレッジの関係

評価データーのパターンに対して、標本データーでn回以上出現するパターンのリコール、プレシジョン、カバレッジを下記に従って求めた。

リコール

$$= \frac{\text{標本データーにn回以上、評価データーに1回以上出現するパターン数}}{\text{評価データーに1回以上出現するパターン数}}$$

プレシジョン

$$= \frac{\text{標本データーにn回以上、評価データーに1回以上出現するパターン数}}{\text{標本データーにn回以上出現するパターン数}}$$

カバレッジ

$$= \frac{\text{標本データーにn回以上出現するパターンの評価データーにおける延べ出現数}}{\text{評価データーのパターンの延べ出現数}}$$

3.3. パターンの出現回数の変動

偏って出現するパターンや出現の頻度が一定であるパターンを抽出するために、コーパス全体から抽出したパターンについて月別の出現頻度の変動係数を算出し、変動係数が高いものと低いものを調査した。変動係数によってパターンの平均出現数と毎月の出現数の違いを見ることができる。また、平均出現数で補正しているので、平均出現数の多いパターンと少ないパターンを同時に比較できる。

4. 評価結果

4.1. 抽出パターン例

コーパス全範囲から1,960,425個のパターンが抽出された。頻度が高かったパターンの例を表5に示す。

例6をみると、もとの文には「空気」を説明する修飾節が含まれているが、パターンには修飾節がなくなっている。節単位のパターン抽出での汎化が成功していることがわかる。

例4をみると、普通名詞の汎化を行なっていないため、パターンの出現頻度が話題に影響を受けている。しかし、抽出されたパターンの自動翻訳への利用を考慮すると、放送ニュースでは事件名を言い換えないため、汎化の判断が難しい。

例7のもとの文をみると、「任期満了に伴う愛媛県の知事選挙は」を「任期満了に伴う愛媛県」で抽出している。これは係り受け解析の誤りが影響している。

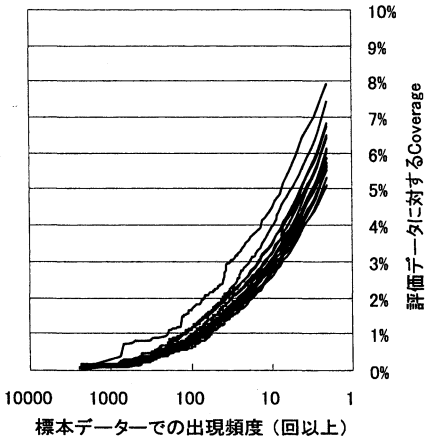
表 5 頻度が高かったパターンの例

パターン	頻度（回）	順位
例1. これを受けて	2467	1
例2. 恐れがあります	846	2
例3. 恐れがあり	814	3
例4. 特別養護老人ホームの建設をめぐる汚職事件	186	42
例5. PNOUN対PNOUNはPNOUNがNUM対NUMで勝ちました	172	48
例6. 市場関係者は空気が強いと話しています	103	101
例7. 任期満了に伴うPNOUN	101	105

4.2. 抽出されたパターンのカバレッジ

図1に35回のクロスバリデーションで標本データーから抽出したパターンのカバレッジを示す。

図1 Coverage (35回のクロスバリデーション)



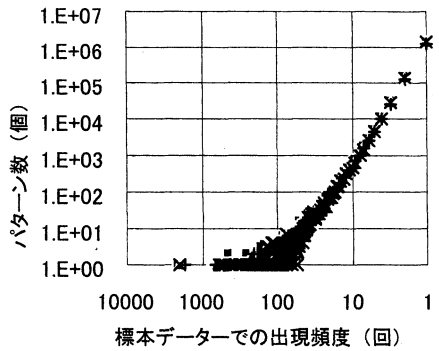
パターンは標本データーで2回以上出現したものを評価した。図1では、35回のうち19回分をプロットした。残りの16回でも同様の傾向が見られた。

カバレッジが比較的高いグラフが1つある。このグラフは標本データーでの出現頻度が2回以上のとき、カバレッジが7.9%となり他と比べて約2%高かった。気象に関する定型的なパターンを含む原稿が多く出現していた1995年7月のグラフである。

4.3. パターンの頻度分布

図2に5回のクロスバリデーションの標本データーから得たパターンの頻度とパターン数の分布図を重ね合わせて示す。標本データーでの出現頻度が1、2回のところで急激にパターン数が増加している。これは、汎化の不足が主な原因と考えられる。

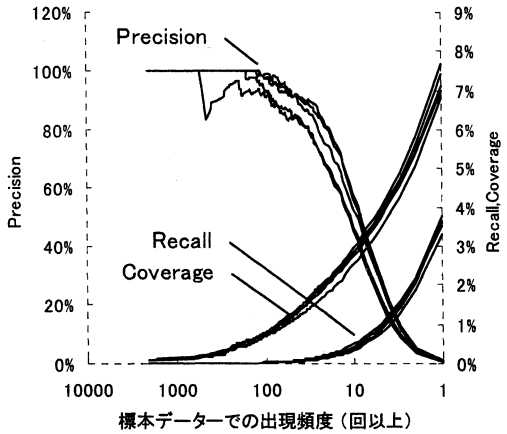
図2 標本データーのパターン数



4.4. リコール、プレシジョン、カバレッジ

5回のクロスバリデーションで求めたリコール、プレシジョン、カバレッジを図3に示す。全体の傾向は同じである。ただし、高頻度のパターンでプレシジョンに変動がある。これは、標本データーでの出現頻度が100以上のパターンの数が極端に少なく、数値が大きく変動しやすいことが主な原因と考えられる。

図3 Precision, Recall, Coverage



4.5. ジャンルの限定

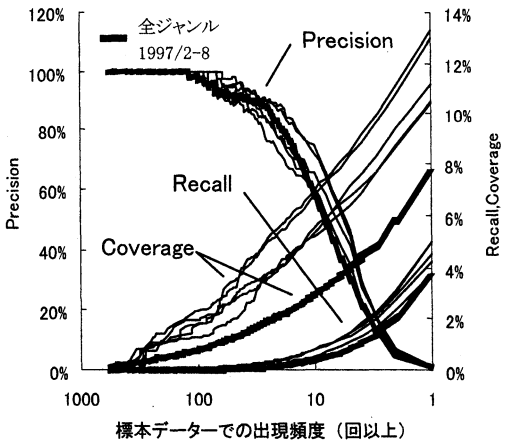
標本データー、評価データーのジャンルを経済に限定して5回のクロスバリデーションでリコール、プレシジョン、カバレッジを求めたので図4に示す。 ここでは比較のために、全ジャンルを対象とした4.4節の結果のうち、1997年2月から8月までを評価データーとするグラフを重ねた。

5回のクロスバリデーションの結果は、全ジャンルを対象とした場合に比べ、相互の違いが出ている。表1のジャンル別の原稿数を見るとデーターの規模が約1/10であり、これが影響していると考えられる。コーパスの規模が若干不十分といえる。

また、カバレッジを全ジャンルを対象とした場合と比較すると約1.5倍から多いところでは2倍以上となっており、高頻度のパターンでもその傾向が続く。表現の多様性の減少を狙ってジャンルを限定した効果が認められる。

7ヶ月ずつとった5つの標本データーの平均を取ると、抽出したパターンは3回出現で約2,800個、2回出現で約14,000個、1回出現では約130,000個であった。経済ジャンルの原稿数がコーパス全体の約1/10であることを考慮し、図2と比較すると、出現数1回のパターン数が相対的に少ないことがわかる。ジャンルを限定しない場合より表現の多様性が減少し効率的にパターンが抽出できている。

図4 経済ニュースでの Recall,Precision,Caverage



4.6. 一時期にしか出現しないパターン

35回のクロスバリデーションで、標本データーから抽出したパターンがカバーできない評価データーのパターン例を表6に示す。

表 6 一時期にしか出現しないパターン

パターン	出現回数（回）	評価データーの範囲（年/月）
例1. NUM時NUM分PNOUN発PNOUN行き	459	1998 / 1
例2. 強い風や雨それに高波に厳重に警戒するよう呼びかけています	87	1997 / 9
例3. 高レベル放射性廃棄物を積んだ輸送船	48	1998 / 3
例4. 地震の回数はNUM回を超えています	48	1997 / 3

例1は大雪で交通が麻痺した月で、文が箇条書きであった。箇条書きは800文字以上の文に多く、これを取り除くためにあらかじめ800文字以上の文を除外していたが、このパターンは文が短かったため除外できなかった。 他の月では同様の箇条書きのパターンは除外できていたため、一時期にしか現れなかったと考えられ

る。このように、月特有のパターンを調査することで、コーパスの内容を均一化する手がかりが得られた。

例2は、特定の季節に出現するパターンと予想されるが、気象情報に関してコーパスが小さいので一時期にしか出現しないパターンとして挙がっていると考えられる。その他には、例3、4のように特定の事件についての話題も多い。

4.7. 毎年特定の月にのみ出現するパターン

毎年特定の月にのみ出現するパターンを表7に示す。

総出現数が9回以上のパターンは537個、3回以上のものまで含むと11,590個であった。3回以上出現するパターンをコーパス全体でカウントすると67663個あった。そのうち約17%のパターンが特定の月に出現していることになる。また、3年で3回しか出現しないパターンの中にも重要なパターンが含まれていることが判かる。

表 7　毎年特定の月に出現するパターン

パターン	総出現回数(回)	出現する月
PNOUNに加盟する全国	56	3月
全国の中小の私鉄やバスの労働組合では大手並みの賃金の引き上げを求めて	50	3月
梅雨前線の活動が引き続き活発なため	50	7月
悪い方向に向かっていると判断されます	24	1月
新年を迎える準備	9	12月
建国記念の日を祝う国民式典	3	2月

4.8. 出現頻度の変動

コーパス全体から抽出したパターンについて、月別の出現頻度の変動係数を算出し、変動が激しいパターンと少ないパターンの調査をした。(表8,表9)

表8　変動が激しいパターン例

激しい雨が降り続く恐れ
PNOUNからPNOUNの広い範囲で雨が降っていて
男は警察官にその場で取り押さえられ
これまでの最高値を更新しました

表9　変動が少ないパターン例

これに関連して
必要があるとしています
身元の確認を急いでいます
NUM点NUM午前の出来高はNUM株でした

変動が激しいパターンには気象関連のパターンが多くみられた。これは、ここに現れている気象関連のパターンが特定の季節や月に出現することが一因であると考えられる。変動が少ないパターンでは、特定の話題が減少し、放送ニュースの特徴的な言いまわしが目立った。また、頻度も多いものが目立った。

5. パターンの抽出・評価方法の考察

5.1. パターンの汎化と話題の特徴

現状の汎化では、低頻度にパターンが集中し、パターンの特徴を観察する上で十分な量の高頻度パターンが確保できていないので、更なる汎化が求められる。

しかし、高頻度パターンである4.1節の例4が示すように、ある話題の決まった表現は、節単位の大きさを持つことがあるので、慎重な汎化が必要である。例えば4..6節で示した一時期にしか出現しないパターンを複数の範囲で比較することで汎化の指針を得られるのではないか。

5.2. 対象ジャンルの絞込み

4.5節では経済にジャンルを限定することで、パターンの抽出精度が上がり、カバレッジが向上することがわかった。しかし、カバレッジの値の揺れが大きくなるなどの問題もみられた。

5.3. パターンの周期性

4.7節では、非常に出現頻度の低いパターンが、毎年同一期間に出現するなど、話題の周期性が確認された。本研究では1年を周期としたが、4年や1週間などさまざまな周期が考えられる。4.6節の例2のように周期的に出現するはずのパターンが一時期にしか出現しないパターンとして抽出されたことからも、特に長い周期の場合は抽出が難しい。

5.4. パターンの出現時期の偏り

4.8節では変動係数を用いてパターンの出現時期の偏りを比べ、変動が多い場合は時期を限定した話題が多く、変動が少ない場合は放送ニュースに特徴的な表現が多くみられることが判った。

6. まとめ

本研究では、ニュース・コーパスからのパターンの抽出とその評価を中心に、ニュース・コーパスの効率的利用につながる分析手法を探った。ジャンル分けによるパターン抽出精度の向上や、標本データーとその評価方法の違いによる、抽出パターンの性質の違いが明らかになった。

今後は、他のジャンルについても分析を行ない、評価の性質の違いから、その結果を系統立てて利用する方法を確立することが課題である。

謝辞

本研究で使用した文節パターン抽出プログラムは、NHK放送技術研究所の江原暉将氏より提供いただいたものを修正して利用した。快くご協力いただき感謝します。

[1]金城由美子, 熊野正, 西脇正通, 柏岡秀紀, 田中英輝. ニュース文のスタイルに関する基礎的調査. 言語処理学会第7回年次大会発表論文集, pp. C-3. 言語処理学会, 2001.