日本語ニュース文パターンの抽出

池崎 健一郎、浦谷 則好、加藤 直人、江原 暉将
NHK放送技術研究所
{ikezaki,uratnai,katonao,eharate}@strl.nhk.or.jp

1　はじめに

我々は日々放送しているニュースを機械翻訳してニュース制作現場の負担を軽減する目的で機械翻訳の研究を行っている。ニュース文は限られた文字数でできるだけ多くの情報を伝えるために、引用や修飾を交えて複雑なものになりやすい。[1]一方で、ニュースには独特の言い回しや定型パターンも多く存在する。日英機械翻訳を行う際、良質な英語翻訳結果を得るためには、日本語パターンを利用することが有効である。そこで、これらを自動的に抽出することを試みた。この結果について報告する。

2　「良い」パターンとは？

ニュース文翻訳にとって、「良い」パターンとはなんであろうか。ニュースは、ある特定の事件などがあった場合、数日にわたってそれらに関する放送を何回も繰り返すために、ただ単純に頻度をとるだけでは、意味のあるパターンは得られない。
たとえば、「通信傍受法案」「日米半導体摩擦」に関するパターンは、それについて議論が白熱している時には連日ニュースに取り上げられるが、いちど問題が沈静化してしまうとほとんどでることはない。つまり、局所的に頻度が非常に高いパターンがあってもそれは将来にわたってパターンとして機械翻訳に役立つとは限らない。

毎日話題がめまぐるしく変わるニュース文で機械翻訳に有効な「良い」パターンとは、ただ頻度が高いだけでなく、株価などの定常的に使われる定型文、ニュース独特の言い回しや慣用表現などであろう。

提案するパターン抽出法では、パターン間の優劣を評価するための関数として**修正頻度**を、パターン自身を抽出するための評価関数として修正相互情報量を用いている。

修正頻度F_pとは、「特定の期間に偏らず、バラバラに出現する表現（パターンp）」に関するパラメータとして**均質性**U_pというものを導入して、実際の出現頻度f_pを次式で修正したものである。

$$F_p = U_p \cdot f_p$$

$$U_p = \frac{\log D_p}{\log n}$$

週明けのきょうの東京株式市場は、政府日銀が円売りドル買いの市場介入に踏み切り、とりあえず円高に歯止めがかかったことなどから、平均株価は大幅に値上がりしました。
きょうの東京市場は、けさ発表された日銀の短観の結果は事前に予想された範囲内だったとして、大きく反応しなかったものの、その後、政府日銀が断続的に円売りドル買いの市場介入に踏み切った結果、先週末以降の円高にとりあえず歯止めがかかったことから、幅広い銘柄に買い注文が膨らみました。
結局、主要銘柄の平均株価、きょうの終値は、先週末より三百八十九円六十七銭高い二万七百二十六円九十九銭でした。
全銘柄の値動きを示すトピックス・東証株価指数は二十六点五一上がって、千七百三十二点四五、一日の出来高は六億九千六百七十八万株でした。

図1．NHK経済ニュースコーパス（一部）

D_p：パターン p の出現する異なり日数

n　：カテゴリ数（総日数）

ここで、均質性U_pとはパターンを抽出するコーパスの中からどれだけ幅広く出現したかを決めるパラメータであり、情報検索で使われる*idf*値

$$idf = \log\frac{n}{D_p}$$

と正反対の性質を持たせる。

一方、修正相互情報量*MMI*とは、相互情報量が頻度の小さいものに対して大きい値を取るという問題を避けるために、次式のように分母を修正した相互情報量である。

$$MMI = \begin{cases} \log\dfrac{f_{xy}}{\dfrac{f_x \cdot f_y}{N}} & \dfrac{f_x \cdot f_y}{N} \geq 1 \text{ のとき} \\ \log\dfrac{f_{xy}}{\dfrac{2}{3}\left(\dfrac{f_x \cdot f_y}{N}\right)^{\frac{2}{3}} + \dfrac{1}{3}} & \dfrac{f_x \cdot f_y}{N} < 1 \text{ のとき} \end{cases}$$

3　パターン抽出

パターン抽出の単位であるが、今回は文節単位で抽出を行うことにした。文字単位・形態素単位に比べ、①得られたパターンが短文であっても比較的意味が把握しやすい、②文字や形態素ごとの計算に比べ計算量が非常に少なくなる、という利点があるからである。

パターンの抽出方法は以下のようなステップから構成される。

Step.1 コーパスに対して、文節認定をする。

例：また、(A) 全銘柄の(B) 値動きを(C) 示す(D) トピックス・東(E)
証株価指数は(F) 二十六点五一 上がって(G) 千七百三十二(H)
点四五で、一日の(I) 出来高は(J) 六億九千六百七十八万株(K)
でした。

Step.2 文節認定の結果に対して、過疎性を回避するために標準化する。

数値→N
人物名→human
場所名→place
組織名→org

に変換し、句点や読点は削除する。

例：また(A) 全銘柄の(B) 値動きを(C) 示す(D) トピックス・東(E)
証株価指数は(F) N点N 上がって(G) N点Nで(H) 一日の(I)
出来高は(J) N株でした(K)

Step.3 修正頻度F_p、相互情報量*MMI*のしきい値をそれぞれ決める。

Step.4 得られたすべての文節(A)(B)(C)...に対して頻度を数える

Step.5 二文節(AB)(BC)...(JK)、三文節(ABC)(BCD)...(IJK)、四文節...と各複数文節に関して頻度f_p、異なり日数D_pを求め、修正頻度F_pを計算する。

Step.6 4で計算した二分節(AB)(BC)(CD)...のFが、しきい値以上の値であれば、記憶しておく。

Step.7 5で出力した(AB)とそれに後接する文節(C)との間の相互情報量*MMI*を求め、しきい値以上なら(ABC)をパターンと認める。

Step.8 (ABC)より長いパターンに対しても、*MI*もしくはF_pがしきい値以下になるまで5と6を繰り返す。

4　実験

前章で述べたパターン抽出法を使って、NHK経済ニュースコーパス1996年～1999年までの4年間を対象に、パターン抽出実験を行った。

図3に1996年のパターン出現頻度f_pと均質性を考慮した頻度F_pのそれぞれ上位10パターンを示す。1996年では「値動きを示すトピックス・

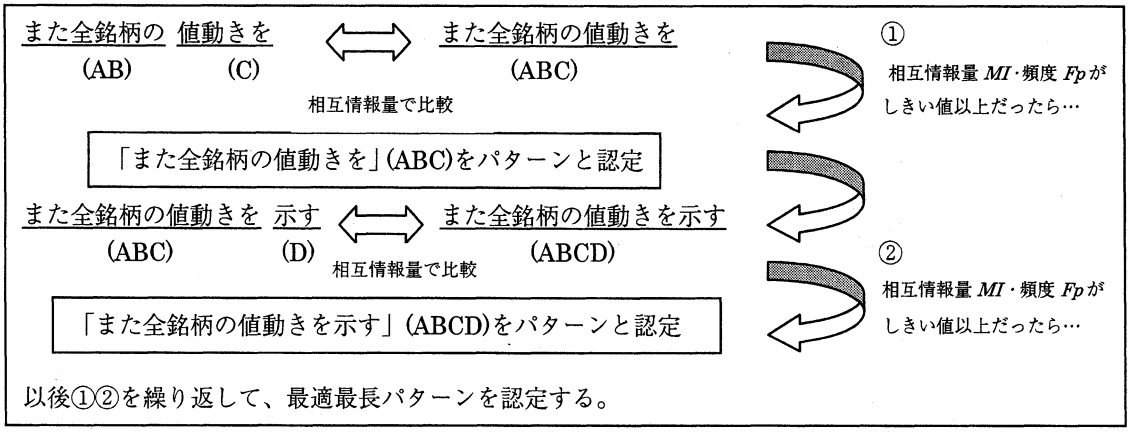

図２：最適最長パターンの認定方法

Fp　*Fp*　*fp*　*Dp*
426.688|460|229|値動きを|示す|トピックス・東証株価指数は
423.301|475|185|一ドルＮ円銭から|銭で|取り引きされています
220.308|271|117|銭円安ドル高の|一ドルＮ円銭から|銭で
206.639|225|217|主要銘柄の|平均株価の|終値は
197.545|243|117|いう|考えを|示しました
197.115|215|215|平均株価|午前の|終値は
196.028|214|214|主要銘柄の|平均株価|午前の|終値は
195.159|236|127|注文が|出て|平均株価は
178.938|223|110|銘柄の|値動きを|示す
178.745|198|198|午前の|出来高は|Ｎ株でした
fp
423.301|475|185|一ドルＮ円銭から|銭で|取り引きされています
426.688|460|229|値動きを|示す|トピックス・東証株価指数は
220.308|271|117|銭円安ドル高の|一ドルＮ円銭から|銭で
197.545|243|117|いう|考えを|示しました
195.159|236|127|注文が|出て|平均株価は
206.639|225|217|主要銘柄の|平均株価の|終値は
178.938|223|110|銘柄の|値動きを|示す
176.989|221|109|銘柄の|値動きを|示す|トピックス・東証株価指数は
169.852|219|94|銭円安ドル高の|一ドルＮ円銭から|銭で|取り引きされています
197.115|215|215|平均株価|午前の|終値は

図３：*Fp* と *fp* 比較(1996 年)

	1996	1997	1998	1999
全文数	22267	25606	31630	33824
パターン認定数	8047	9912	11034	11356
総日数	350	365	361	361
全文節数	417065	499971	609562	606109
カバー文節数	85191	113288	131837	127793
カバー率	20%	22%	21%	21%

図４：パターンカバー率の比較　(*fp*：3 以上　*Dp*：3 以上　*MMI*：2 以上)

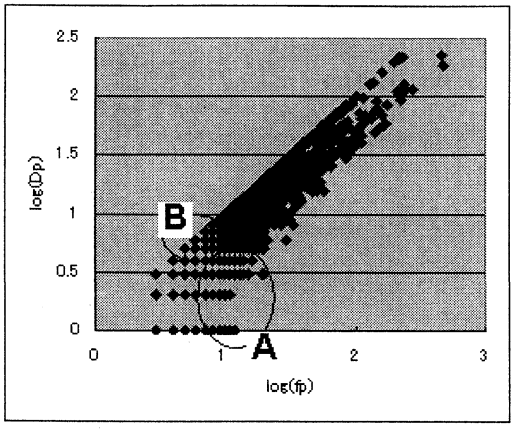

図5：*log(fp)*と *log(Dp)*の散布図

東証株価指数は」がパターンのトップ項目に挙げられ、22267文中パターンとして認識された回数は460回、出現日数229日となっている。他に「主要銘柄の平均株価の終値は」「平均株価午前の終値は」が上位に抽出されており、平日必ず放送される円相場や株式相場がほぼ定型パターンとして抽出されている。また、均質性を考慮した結果、両者の順位はかなり入れ替わる。例えば、「銘柄の値動きを示すトピックス・東証株価指数は」はパターン頻度に比べ出現日数が低いため F_p 値が低くなり、上位に出現しなくなる。

また、図4にパターンカバー率を示す。ここでパターンカバー率とは、抽出されたパターンで全文をどれだけカバーできているかを表している値である。図4を見ると提案手法を用いることにより、全文の約20%程度のパターン化を抽出している。

また図5は $log(f_p)$ と $log(D_p)$ の散布図である。ただし、D_p のしきい値を1に変更している。図5上のA付近は、f_p に比べて D_p が低いところである。Aでは「発展途上国に対する援助の在り方それに貿易と投資の」「収穫から一年以上経っている」など、局所的に偏在するデータを発見できる。これに対してB付近はAよりも f_p が小さいが、D_p は高いので、定常的なパターンを発見しやすい。例えば「先月首都圏で発表された新築マンションは」「発表した機械受注統計によりますと主な機械メーカーが」など毎月報告されるパターンとして抽出されている。

5　考察

本稿では単語の出現頻度に均質性を考慮した値 F_p を使用して、パターンそれぞれの評価基準としている。これにより、定常的なパターンを抽出することができた。

しかしながら、パターン化の結果は20%と期待した程ではなく、このままでは充分とはいえない。これについて、今後は①標準化作業を拡大する、②今回は、連続する文節だけの評価だったが、係り受け情報も使用する。ことによって離れた共起も見るなどにより、いっそうの被覆率の向上を図り、ニュース文のパターン翻訳に役立てる予定である。

【参考文献】

[1]NHK放送文化研究所.放送基本語彙調査,1989.
[2]Carroll,J.B.,Davis,P.,andRichman,B.The American Heritage Word Frequency Book. Boston,MA:Houghton Mifflin,1971
[3]Alex Chengyu Fang and Mark Huckvale. Out-of-Vocabulary Rate Reduction through Dispersion-Based Lexicon Acquisition:LLC,Vol 15,No.3,2000