7B-6

テキスト・データベースからの慣用表現の自動抽出

北　研二　　森元　逞

ＡＴＲ自動翻訳電話研究所

１．はじめに

自然言語処理の分野では、従来から慣用表現の重要性が指摘されており、慣用表現を網羅的に収集する試みもいくつかの研究機関で行われてきた。しかし、慣用表現の抽出は、これまではすべて人手で行われており、膨大な手間と時間を要した。また、慣用表現抽出の基準も曖昧であった。

本稿では、慣用表現抽出の基準を定式化することを試みる。また、提案する方法を用いて、実際のテキスト・データベースから慣用表現を抽出した結果を示す。

電話会話などでは、慣用表現が多用されるため、これらの文の解析等へ本方法を適用することにより、大幅な効率向上が期待できる。

２．基本的な考え方

基本的な考えは、テキスト中に多く出現する単語列を見つけ出すということである。しかし、単に単語列といっても、２つの単語から成る列もあるし、３つの単語から成る列もあり、単純に単語列の出現回数で比較するわけにはいかない。

例えば、あるテキスト中に、"in spite" という単語列が１１０回、"in spite of"という単語列が１００回、"in spite of XXX" という単語列が２０回出現したとする。この場合、"in spite" が最も出現回数が多いからといって、単純に "in spite" を慣用表現とするのには問題がある。なぜならば、"in spite" という用いられ方において、ほとんどの場合はその後に "of" を伴って用いられているからである。いまの場合は、"in spite" は無視して "in spite of" を慣用表現としなければならない。

このように、テキスト中から慣用表現を抽出する際には、単語列の出現頻度と、単語列が生成される系列（上の例の場合、"in spite" -> "in spite of" -> "in spite of XXX" ... という系列）間の出現頻度の差を同時に考慮する必要がある。

３．慣用表現抽出の基準

最初に、いくつかの表記法を導入する。αを単語列とするとき、次のような表記法を用いる。

$|\alpha|$ ・・・ 単語列の長さ（いくつの単語から成る単語列であるかを示す）.

$n(\alpha)$ ・・・ 単語列の出現回数.

単語列の比較に用いる尺度に、我々は以下の式の値を用いる。すなわち、下の式の値の大きい単語列を慣用表現とみなし、テキスト・データベースから抽出する。

$$K(\alpha) = (|\alpha| - 1) \times n(\alpha)$$

まず、$K(\alpha)$の意味付けを行う。

いま、テキスト中にｎ個の単語から成る慣用表現αがあったとする。もし、αをｎ個の単語から成るものとして処理すれば、ｎに比例するだけの仕事量が必要である。しかし、αを１つの表現として処理すれば、仕事量は１でよい。すなわち、（ｎ－１）分の仕事量が節約されたことになる。αが、テキスト中にｍ回出現するならば、

$$(n-1) \times m$$

分の仕事量が節約される。これが$K(\alpha)$の意味付けである。

次に、同一の単語列生成系列中の２つの単語列の扱いについて説明する。

αとβが同一の単語列生成系列に属し、しかも$|\alpha| < |\beta|$であったとする（すなわち、αはβの部分単語列である）。このとき、当然のことながら、

$$n(\alpha) \geq n(\beta)$$

である。αとβの両方を慣用表現としてテキストを処理する場合を考える。βはテキスト処理の際にｎ（β）回参照されるが、αについて考えると、αの出現回数ｎ（α）のうちｎ（β）回についてはβが参照されているので、純粋にαが参照されるのは、

$$n(\alpha) - n(\beta)$$

回だけである。従って、αとβの両方を慣用表現として採用する場合には、αの持つＫの値を

$$(|\alpha| - 1) \times (n(\alpha) - n(\beta))$$

に変更する必要がある。

Automatic Idiom Extraction from Text Database
Kenji KITA, Tuyoshi Morimoto
ATR Interpreting Telephony Research Laboratories

4．インプリメンテーション

テキスト中から慣用表現を抽出するためには、まず、テキストを走査し、各単語列の出現回数を調べ、次に、同一の単語列生成系列に属する単語列のKの値を動的に変更しながら、Kの値の大きい方から単語列を順次取ってくる必要がある。

後者の操作を厳密な意味で実現しようとすると、計算量が爆発的に増大する。なぜならば、単語列を1つ取ってくる度に、これまでに取られた単語列のすべてに対して同一の生成系列に属するか否かを調べて、Kの値を再計算しなおさなければならないからである。

例で説明する。いま、単語A，B，C，D，E・・・に対して、単語列の出現回数が図1のようになったとする。

このとき、まずKの値の最も大きいAB（K＝100）が取られる。次に、2番目にKの値の大きいABCD（K＝90）が取られるが、既に取られているABは、ABCDの部分単語列となっているので、ABの持つKの値は70に修正される（再計算1）。このあと、ABCが取られ、再度ABの持つKの値が再計算される（再計算2）。同様に、ABEが取られた時点でも、ABに対して再計算が行われる（再計算3）。

上の例は、比較的単純な例であるが、実際には単語列は1方向だけではなく、両方向に成長するため、より複雑な計算が必要となる。

我々は、計算量の爆発を防ぐために、再計算の操作を近似的に行うこととした。この方法は、以下のようなものである。

(1) Kの値の再計算は生成系列中で隣合った単語列どうしに限る。上の例の場合、ABとABCDは生成系列中で隣合っていないので、再計算1は行われない。

同一の生成系列中で単語列の長さが2以上離れているものについての再計算は、隣合った単語列どうしの再計算に帰着される。従って、生成系列中の再計算は原則的には隣合った単語列どうしでいいが、Kの値の大きい順で取ってきたときに必ずしも隣合った単語列が順に取ってこられるとは限らない。しかし、順に取ってくるかぎり、いずれは生成系列中で隣合った単語列が現われ、その際に再計算をすれば十分であるという理由に基づいている。

(2) 再計算は1度だけに限る。上の例の場合、ABCが取ってこられた時点で、ABの再計算が行われているので、ABEが取ってこられてもABの再計算は行わない。つまり、再計算3は行われない。

これは、単語列をKの値の大きい順に取ってきているため、後からされる再計算ほどKの値に与える影響が少ないという理由に基づいている。

5．実験結果

上で提案した手法の有効性を確認する実験を行った。実験には、ATR自動翻訳電話研究所が収集した言語データベース用の基礎資料のうち、電話会話を対象としたものを用いた（総文数 2758，総単語数 32339，異り単語列数 156598）。

単語列の生成系列は10単語から成るものまでを扱った。また、単語列のうちKの値の大きいもの0．5％についてのみKの再計算を行った。表1に慣用表現として抽出されたものの一部を示す。

これを見ると、比較的妥当と思えるものが抽出されており、本レポートで提案した手法の有効性がわかる。

K: <N-gram> Occurrences: Word-sequence

460: <3>370: でしょ,う,か
348: <4>140: ん,でしょ,う,か
326: <3>221: ん,です,けれども
322: <2>411: の,方
255: <4> 85: あ,そう,です,か
254: <3>212: そう,です,か
222: <4> 74: はい,わかり,まし,た
192: <3>123: と,いう,こと
160: <2>381: ん,です
156: <2>221: です,ね
150: <4> 65: ます,でしょ,う,か
144: <4> 48: ああ,そう,です,か
143: <2>176: し,て
142: <3> 92: と,思い,ます
141: <4> 58: な,ん,です,けれども
140: <3> 89: の,方,に
132: <3>121: て,おり,ます
128: <3> 64: の,方,は
126: <3>137: わかり,まし,た
120: <5> 45: に,なっ,て,おり,ます

表1．電話会話からの慣用表現抽出結果

6．おわりに

慣用表現抽出の基準を定式化し、テキスト・データベースから慣用表現を効率的に収集する方法について述べた。また、提案した方法の有効性を、実際のテキスト・データベースを用いて検証した。

単語列	出現回数	最初	再計算1	再計算2	再計算3	・・・
AB	100	100	70	60	40	
ABCD	30	90	90	90	90	
ABC	40	80	20	20	20	
ABE	20	40	40	40	40	
・・・	・・	・・	・・	・・	・・	

図1．再計算の様子