7L－4

ギャップのある n-gram を用いたテキストコーパスの解析

國吉 芳夫　中西 正和
慶應義塾大学 理工学研究科 計算機科学専攻

1. はじめに

テキストコーパスの共起統計量に基づく解析では、隣接する2または3単語の列の頻度であるbigramやtrigramが使用されることが多い。これまで4より大きいnに対するn-gramはあまり用いられてこなかったが、長尾等の方法[1]は大きいnに対するn-gramを効率良く求める方法を与えた。

本稿では、任意のnに関するn-gramについて各単語の間に任意の単語の挿入を許した、ギャップのあるn-gramを提案し、その特別な形である、ギャップが1つだけの場合について行った簡単な実験の結果について報告する。

2. n-gram統計量を求めるためのアルゴリズム

長尾等のアルゴリズムは極めて単純である。

まず、対象となるコーパスの各語を指すポインタの配列を作成し、このポインタが指しているところから始まる単語列に関してソートする。結果の配列を上から順に見ていけば、コーパスに含まれているすべてのn-gramの頻度を1回の走査で求められる[1]。

3. ギャップのある n-gram

ギャップのあるn-gramとは、n-gramを構成する各単語の間に任意の単語が0個以上入ることを許したものである。これにより、隣接する単語に基づくn-gramモデルでは抽出することが不可能であった、距離的に離れて共起する単語列を捕えることができるようになる。距離的に離れた単語列の抽出ができるようになれば、より大きな統語的構造や、格フレームに関する制約を直接取り出せるようになると考えられる。

3.1 定義

ギャップのあるn-gramは次のような長さnの単語列のコーパス中における頻度を各$n(\geq 2)$について求めたものである。ただし、$d(w_i, w_j)$をw_iとw_jとの距離(ギャップの長さ)とする。距離は隣接する単語のとき0、1単語間に入るごとに1増えるものとし、負の距離は考えない。

$$w_1 w_2 \cdots w_n \text{ where } d(w_i, w_{i+1}) \geq 0 \qquad (1)$$

一般には、ギャップのあるn-gramはギャップの長さや位置が異なるものを区別して考える。しかし、ギャップの長さや位置などに関して同一視するための制約を持ち込むことにより、異なる性質のn-gramを考えることができる。

3.2 ギャップの配置を固定した n-gram

各ギャップの位置が同じであれば、ギャップの長さが違っても同一のn-gramとみなすようにすると、ギャップのないn-gramをいくつか並べたものを表すことになる。

また、これの簡単な場合である、ギャップの個数が1であるものをbi-n-gramと呼ぶ。

3.3 ギャップ数が1で長さも固定した n-gram

ギャップのあるn-gramの頻度の計算としては、ギャップのないn-gramの表を用いて先頭の部分が一致するものについてそれぞれ数える方法があるが、これでは最初のギャップの前にある単語列の長さが短いときには非常に効率が悪い。

そこで、$d(w_1, w_2) > 0$であるbi-n-gramのうち、頻度の高いものについてあらかじめ求めておくと良

Yoshio KUNIYOSHI　(yoshio@nak.math.keio.ac.jp)
Masakazu NAKANISHI
Department of Computer Science, Graduate School of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223, Japan

[1]ポインタを逆にたどることも考えれば、ここで作成した配列は文脈無制限のKWIC表になっている。

い。それには、各 d について i 番目以降の単語を d 個スキップしたもの (ギャップ長が d) に対して n-gram を求めれば良く、それは、節 2. で述べたアルゴリズムにおけるソーティングで使用する比較関数を変更するだけで良い。

4. bi-n-gram によるコーパスの解析

4.1 独立な単語列

bi-n-gram を構成する 2 つのギャップのない n-gram として独立な単語列を取ることを考える。独立な単語列とは、周囲の環境とは独立に高い頻度でコーパスに現れる単語列のことである。

独立な単語列はその左右に現れる単語が多様であると考えられることから、左右の単語のエントロピーを尺度として抽出することができる [1]。

右側のエントロピーは n-gram 表から簡単に求められる。左側のエントロピーを求めるには、コーパスを逆順にしたものに関して作成した n-gram 表を用いることができる[2]。

4.2 bi-n-gram の評価値

bi-n-gram に評価値を与えることを考える。bi-n-gram はギャップのない 2 つの n-gram で構成されているから、一方の n-gram が現れたときのもう一方の n-gram の現れ易さを尺度とするのが良い。これには、相互情報量の考え方が使える。このとき、評価値の相対的な比較だけを目的とするならば、次の関数を用いれば十分である。

$$\frac{c(left, right)}{c(left) \cdot c(right)} \qquad (2)$$

ここで、$c(x)$ はギャップのない n-gram x の頻度、$c(x,y)$ はギャップのない n-gram x, y が一定の距離内に同時に現れる頻度、*left*, *right*はそれぞれギャップの左側の n-gram と右側の n-gram である。

4.3 句の認識への応用

高い評価値を与える bi-n-gram を構成する 2 つの n-gram は、互いを関係付ける何らかの統語的な規則を反映している可能性が高い。したがって、ギャップに入る単語列を含めて、一つの句を形成している可能性が高い。そこで、上位の bi-n-gram に関してコーパスを検索し、ギャップに入る単語列を、その bi-n-gram を構成する 2 つの n-gram でカテゴライズすることができると考えられる。

[2] ソートする際の文字列の比較においてポインタを逆にたどるだけで良いので、逆順のコーパスは必要ない

5. 実験

bi-n-gram の抽出実験を行った。抽出の評価値としては式 2 を用い、テキストコーパスのすべての文に対して独立な単語列からなる bi-n-gram を長さが 3 以上、ギャップ長 20 以下のものについて上位 10000 個求めた。独立な単語列は、各文毎に、文中の各語について左右に一語ずつ延ばしながら左右のエントロピーを求め、それが最大値を与える場所で切ることによって取り出した。

bi-n-gram の抽出及び n-gram 表の作成に使用したのは語彙数 62000 語、文数 54000 文、総語数 112 万語の英語テキストコーパスである。

6. 結果と考察

評価関数の性質により、頻度が小さいものが上位に現れやすい。特に、一方が頻出する n-gram で、もう一方が固有名詞など稀にしか使われないものというペアが非常に多く抽出されてしまった。逆に、"as ... as possible" のような良く使われる言回しでも、as が頻出語であるために評価値が小さくなり抽出できなかった。統語的に意味のあるペアを抽出するためには評価関数の再検討が必要である。

ギャップの最大長を 20 としたが、この値は英語のコーパスでは大き過ぎて統語的に関係のない組が上位に現れる原因となった。

参考文献

[1] Makoto Nagao, Shinsuke Mori. 1994. A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese. In *COLING 94*, pp. 611–615.