3R－6

最大エントロピー法を用いた離散型共起表現の抽出

廣鳴 伸章　佐藤 健吾　中西 正和
慶應義塾大学大学院 理工学研究科 計算機科学専攻

1. はじめに

最近、統計的自然言語処理に関する研究が盛んに行われている。その中で頻繁に利用される共起表現には、連続した位置に共起する連鎖型の共起表現と、離れた位置に共起する離散型の共起表現の2つに大別されるが、離散型の共起表現は分野によって使用される表現に違いがあるため、手作業で抽出することは困難となっている。

そこで本論文では、離散型の共起表現を自動抽出することを目的とする。そのための手法としては、従来のような連続型の共起表現を組み合わせる手法[1, 2]ではなく、最大エントロピー法をヒューリスティックスとして用いる手法を提案する。

2. ギャップのある n-gram

ギャップのある n-gram は、任意の数の任意の単語で中断されることを許した n 単語の列を表すモデルであり、以下のように表せる。

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

ここで $d(w_i, w_j)$ は単語 w_i と w_j との間に入る任意の単語の数を表す。このモデルは、$i = 0, \cdots, n-1$ に対して $d(w_i, w_j) = 0$ のとき連続型、そうでないときには離散型の共起表現を表すものである。

3. 最大エントロピー法

最大エントロピー法を自然言語処理に応用した研究として、bigram の推定に応用した研究[3]、n-gram の推定に応用した研究[4]などがある。後者は n-gram 確率を最大エントロピー法を用いて bigram 確率で補完するというものであるが、本論文ではこの手法を改良し、ギャップのある n-gram の確率について補完を行なう。

3.1 最大エントロピー法の適用

ある言語 $\mathcal{L}$ の単語の集合を $\mathcal{W}$ とする。$\mathcal{L}$ の中で $\mathcal{W}$ の要素が n 個出現する事象を表す確率変数を $W_1, \cdots, W_n$ とする。このとき、$w_i \in \mathcal{W}, i = 1, \cdots, n$ に対して単語列 $ws \in \mathcal{L}$ が (1) 式の形で表せるとき、確率 $p(W_1 = w_1, \cdots, W_n = w_n)$ が考えられる。これをギャップのある n-gram の確率と呼ぶことにし、$p(ws)$ と書くことにする。このギャップのある n-gram の確率によって確率変数 $W_1, \cdots, W_n$ の同時分布が得られる。また、$i = 1, \cdots, n-1; j = i+1, \cdots, n$ に対して、$p(W_i = w_i, W_j = w_j)$ が考えられる。これは bigram 確率と呼ばれ、$p(w_i, w_j)$ と書かれる。これによって W_i, W_j に関する2次の周辺分布が得られるが、周辺分布は同時分布から周辺和をとることによって求めることができる。すなわち、

$$p(w_i, w_j) = \sum_{w_1 \in \mathcal{W}} \cdots \sum_{w_{i-1} \in \mathcal{W}} \sum_{w_{i+1} \in \mathcal{W}} \cdots \sum_{w_{j-1} \in \mathcal{W}} \sum_{w_{j+1} \in \mathcal{W}} \cdots \sum_{w_n \in \mathcal{W}} p(ws) \tag{2}$$

によって計算できる。

しかし、逆に bigram 確率からギャップのある n-gram の確率を求めることは、周辺分布から同時分布を求めることであり、解は一意的には定まらない。そこで、最大エントロピー法を適用することを考える。エントロピー

$$H(p) = -\sum_{w_1 \in \mathcal{W}} \cdots \sum_{w_n \in \mathcal{W}} p(ws) \log p(ws) \tag{3}$$

を最大とする p をギャップのある n-gram の確率の推定値 $\hat{p}(ws)$ とするものである。つまり

$$\hat{p}(ws) = \arg\max_p H(p) \tag{4}$$

とする。ここで p は (2) 式を満たさなければならない。このような制約付きの最大値を求める問題は Lagrange の未定係数法を用いることによって解くこと

Extracting Interrupted Collocations Using Maximum Entropy Method
Nobuaki HIROSHIMA Kengo SATO Masakazu NAKANISHI
Department of Computer Science, Graduate School of Science and Technology, Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan

ができ、その解は以下のようになる。

$$\hat{p}(ws) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} h(w_i, w_j) \qquad (5)$$

ここで、h は未定の関数であり、比例反復法を用いて以下のように求めることができる。$h(w_i, w_j)$ の $r+1$ 回目の反復値は、r 回目の反復値を用いて、

$$h_{r+1}(w_i, w_j) = \frac{p(w_i, w_j)}{p_r(w_i, w_j)} h_r(w_i, w_j) \qquad (6)$$

となる。ここで、$p_r(w_i, w_j)$ は $h_r(w_i, w_j)$ を (5) 式に代入して計算した $\hat{p}(ws)$ を $p(ws)$ として (2) 式に代入することによって得られるものである。

4. 本論文の方法

次の手順に従って離散型の共起表現の抽出を行う。

1. コーパスから各 bigram (w_i, w_j) の度数 m_{w_i,w_j} を調べ、周辺分布 $p(w_i, w_j)$ を求める。
2. (6) 式に初期値 $h_0(w_i, w_j)$ を与えて比例反復法を実行し、$\hat{p}(ws)$ を求める。
3. 得られた $\hat{p}(ws)$ の値をもとに順位づけし、順位の高いものを離散型の共起表現として抽出する。

5. 実験および評価

5.1 実験方法

本論文の方法の効果を検証するため、ETL コーパス (英文コーパス、延べ単語数 26 万語、122 万文) を対象に、3 単語からなる離散型の共起表現の抽出実験を行った。コーパス中に出現する大文字は、WordNET の辞書を用いてすべて小文字に直した。また、動詞はすべて原形に戻した。bigram の頻度の最大値および最小値については、先だって調べた bigram の度数分布に基づき、最小値 9、最大値 30 として実験を行った。その結果を表 1 に示す。

表 1: 実験結果

順位	trigram 確率	trigram		
1	2.424413e-05	most	to	people
2	2.244996e-05	speak	to	english
3	2.061977e-05	to	traffic	accident
4	1.926786e-05	speak	the	english
5	1.903371e-05	come	be	again
6	1.679438e-05	to	solve	problem
7	1.664101e-05	he	traffic	accident
8	1.358732e-05	i	solve	problem
9	1.251509e-05	of	leave	home
10	1.169151e-05	to	cut	down
11	1.021918e-05	learn	the	english
12	1.009269e-05	the	speak	english
13	9.982744e-06	cut	down	expense
14	9.674344e-06	learn	to	english
15	9.565778e-06	cut	to	down

5.2 評価

この結果を見ると、日常生活で用いられる慣用的な文の一部と見られる共起表現が多く抽出されている。このような共起表現が抽出された背景には、次のようなことが考えられる。

まず第一に、本実験で用いたコーパスは、新聞記事などのように独特の言い回しを多く含むような文ではなく、日常生活をモチーフにした文が大半を占めているということである。そのため、上位の trigram に含まれる名詞に着目すると、日常生活に密着した単語が出現している。それらの単語のコーパス中での出現頻度はどれも非常に大きい。次に、3 単語からなる離散型の共起表現には "as ... as possible" のような重要な表現は少ないということも挙げられる。連続型の共起表現と比較して、その組み合わせは多彩であるにも関わらず、それらの重要な表現は少ないのであるから、それが上位 15 位以内に含まれることは極めて稀であると言ってよい。

これらを考えあわせると、今回抽出された共起表現は妥当なものであるとしてよい。

6. おわりに

最大エントロピー法を用いて離散型の共起表現を抽出する方法を提案し、その方法が離散型の共起表現の抽出に有効に働くことを示した。

参考文献

[1] 池原悟, 白井諭, 河岡司. "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". 情報処理学会論文誌, Vol.36, No.11, pp.2584-2595, 1995.

[2] 國吉芳夫, 中西正和. "ギャップのある n-gram による言回しの抽出", 情報処理学会自然言語処理研究会, 1997.

[3] 白井清昭, 乾健太郎, 徳永健伸, 田中穂積. "最大エントロピーを用いた単語 bigram の推定". 自然言語処理, Vol.116, No.4, pp.21-28, 1996.

[4] 江原暉将. "最大エントロピーを用いて n グラム確率をバイグラム確率で補完する方法". 言語処理学会第 2 回年次大会発表論文集, pp.369-372, 1996.