d-bigram を用いた自然言語文評価に関する実験

1H－2

佐野 智久, 堤 純也, 孫 大江, 延澤 志保, 佐藤 健吾, 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

“統計情報”を用いた自然言語処理では、“コーパス”と呼ばれる大量の例文のサンプル集をもとに解析を行なわれてきている。本研究では、品詞や構文の情報などを付加していない文から成るコーパスから得られる d-bigram 情報を用いて自然言語文を評価することに関して考察することを目的とする。

2. d-bigram

統計情報を用いた自然言語処理では、bigram や trigram などの n-gram と呼ばれる確率モデルが用いられることが多い。n-gram とは、ある n 個の出現順序を考慮した単語列 $w_1, w_2, ..., w_n$ の同時生成確率モデルである。$n=2$ のもの（2 単語が連続して現れる確率のモデル）を bigram、$n=3$ のものを trigram と呼ぶ。P. Brown は、trigram を用いて文の評価を行なった [1]。

bigram に n-gram と 相互情報量の考え方に基づき、距離の概念を採り入れた確率モデルが d-bigram である [2][3]。d-bigram は 2 つの単語が距離 d で現れる確率を統計情報として保持している。すなわち、連続してはいないが関係のある単語同士の情報も保持できるという特徴がある。

3. 文の評価式

d-bigram を用いた 2 単語間の相互情報量として、

$$MI(x_i, x_{i+d}, d) = \log_2 \frac{P(x_i, x_{i+d}, d)}{P(x_i)P(x_{i+d})} \quad (1)$$

x_i : 入力列中の i 番目の単語
d : 2 単語間の距離
$P(x)$: 単語 x が現れる確率
$P(x, y, d)$: 単語 x, y が距離 d で現れる確率

を用いる。次に、文に対する評価値は文の中の全ての単語の組に対してのそれぞれの相互情報量の和を値とする [2][3]。

$$I(W) = \sum_{d=1}^{m} \sum_{i=0}^{n-d-1} MI_d(x_i, x_{i+d}, d) \quad (2)$$

W : 入力文
n : 入力文の単語数
m : d-bigram の最大距離

4. 実験

4.1 d-bigram による相互情報量に対する重み付け

d-bigram を用いた相互情報量は距離 d に依存するが、距離が遠くなるほどこの相互情報量に対するノイズが大きくなると考えられる。そこで文の評価値を計算する際に、2 単語間の距離が遠くなるほど評価値に与える影響を減らすように、相互情報量に対しての重み付けを次のようにした。

$$I_d(W) = \sum_{d=1}^{m} \sum_{i=0}^{n-d-1} \frac{MI(x_i, x_{i+d}, d)}{g(d)} \quad (3)$$

$g(d)$ は d-bigram を用いた相互情報量に対する重み付けで次のようなものが考えられる。

$g(d) = 1$ $(const)$
$g(d) = d$ $(linear)$
$g(d) = d^2$ $(square)$
$g(d) = e^d$ $(exponent)$

相互情報量をある定数で割ったもの、距離で割ったもの、距離の二乗で割ったもの、距離の指数で割ったものなどが考えられる。

ここでは、ランダムに選んだ 6 以上 10 以下の単語からなる文をそれぞれ 50 文ずつの 500 文用意した。これを用いて、それぞれの重み付けに対して実験した。

4.2 窓かけ

ある 2 単語が現れる時、その距離 d を一意に決めてしまうのは無理がある。2 単語が距離 d の近傍に現

An Experiment on Evaluating Natural Language Sentences Using D-bigram
Tomohisa SANO, Junya TSUTSUMI, Da Jiang SUN, Shiho NOBESAWA, Kengo SATO, Masakazu NAKANISHI
Department of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

れるというように考えるのが自然である。そこで、式(1)の相互情報量を窓かけして次のように拡張した。

$$MI_w(x_i, x_{i+d}, d) = \sum_{w=w_{min}}^{w_{max}} \log_2 \frac{MI(x_i, x_{i+d+w}, d+w)}{f(w)} \quad for \quad 0 \leq i+d+w < n \tag{4}$$

w : 窓の大きさ
w_{max} : 窓の範囲の上限
w_{min} : 窓の範囲の下限

$f(w)$ は窓かけに対する重み付けであり、相互情報量の時と同様に次のようなものが考えられる。

$f(w) = 1$ $(const)$
$f(w) = |w| + 1$ $(linear)$
$f(w) = (|w| + 1)^2$ $(square)$
$f(w) = e^{|w|+1}$ $(exponent)$

4.1と同じ例文を用いて、それぞれの窓の大きさ、重み付けに対して実験した。

5. 結果

入力した文と全く同じ文が得られた場合を正解とした場合の結果は以下のようになった。

表1は、4.1で述べた $g(d)$ の検証をした結果である。4.2による効果を無くすために、式4中の w を0にしている。

表1: 評価式における $g(d)$ の検討

$g(d)$	1位	20位以内
const	28.82	62.35
linear	28.61	60.83
square	25.64	55.95
exponent	25.28	52.27

以上のように、

$$const > linear > square > exponent$$

の順に結果が良くなっている。

表2、3は、4.2で述べた $f(w)$ の検証をした結果である。4.1による効果を無くすために、式3中の $g(d)$ は $square$ に固定する。2は正解文が1位に得られた場合、3は正解文が20位までに得られた場合である。

表2: 窓かけにおける $f(w)$ の検証 (1位)

$f(w)$	0	1	2
const	25.64	34.00	40.70
linear	25.64	41.70	34.80
square	25.64	28.91	32.01
exponent	25.64	35.20	35.20

表3: 窓かけにおける $f(w)$ の検証 (20位以内)

$f(w)$	0	1	2
const	55.95	75.20	81.40
linear	55.95	82.41	76.00
square	55.95	62.04	67.32
exponent	55.95	74.80	75.20

以上のように、ウィンドウサイズが0の場合には窓かけの影響は現れないため、正答率はすべて同一となる。ウィンドウサイズが1の場合 $f(w)$ は、

$$linear > const > exponent > square$$

の順に良い性質を示す。しかしながら、ウィンドウサイズが2の場合 $f(w)$ は、

$$const > linear > exponent > square$$

の順に良い性質を示す。

6. 評価

前節の結果のように、$g(d)$、$f(w)$ のそれぞれの重み付けに対する影響はさほど大きなものではない。しかしながら、評価値、窓かけのいずれの場合においても $const$ が良い結果を得ることは興味深い。

今回の実験はそれぞれのケースについて300文程度の試験を行なったが、試験データの偏りが見られ、それが結果に影響を与えている可能性がある。今後、無作為データに対する大量試験を行なうことで、より中立な結果を導くことが可能であろう。また、今回のような大局的な検証だけでなく、試験結果の内容について特徴的なパターンを検証することは文評価式の信頼性を上げるために重要な課題であると考える。

参考文献

[1] Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R. and Roossin, A. A Statistical Approach to Language Translation. *Coling*, pages 79–85, 1990.

[2] 堤 純也, 新田 朋晃, 小野 孝太郎, and 延澤 志保. 統計情報を用いた多言語間機械翻訳システム. 人工知能学会研究会, pages 7–12, 1993.

[3] Tsutsumi, J., Nitta, T., Ono, K., Nobesawa, S. and Nakanishi, M. Multi-lingual Machine Translation Based on Statistical Information. *Qualico*, pages 147–152, 1994.