自然言語文評価における d-bigram 情報の活用方法に関する実験

佐野 智久
堤 純也, 孫 大江, 延澤 志保, 佐藤 健吾, 大森 久美子, 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からなるサンプル集をもとに解析が行なわれている。現在ではコーパスに品詞や構文の情報などを付加したものを用いていることが多い。

本研究では、これら情報を付加していない文からなるコーパスから抽出された d-bigram (距離の概念を採り入れた bigram) 情報を用いている。本稿では、d-bigram を用いた相互情報量を利用して与えられた自然言語文がコーパスの中でどの程度 “ありそうな文” であるかを評価することによって、d-bigram 情報の最も効果的な活用方法に関して考察し、その結果を報告する。

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。

2.1 n-gram

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ている [1]。

2.2 d-bigram

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。

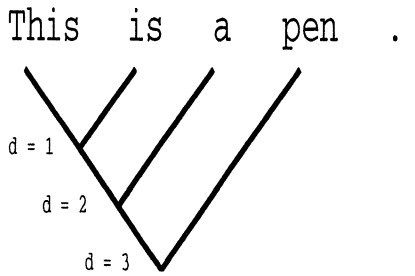

図 1: d-bigram

3 d-bigram を用いた相互情報量

3.1 相互情報量

一般に、事象 a, b の間の相互情報量とは次のように定義される。

$$MI(a,b) = \log_2 \frac{P(a|b)}{P(a)} = \log_2 \frac{P(a,b)}{P(a)P(b)} \quad (1)$$

3.2 d-bigram を用いた 2 単語間の相互情報量

ここでは相互情報量を距離の概念を採り入れた d-bigram を用いて拡張する。

事象を自然言語文における単語の出現とすると、ある 2 単語が距離 d で出現した時の相互情報量となる。この d-bigram を用いた 2 単語間の相互情報量は次のように定義される。

$$MI_d(x_i, x_j, d) = \log_2 \frac{P(x_i, x_j, d)}{P(x_i)P(x_j)} \quad (2)$$

An Experiment on Good Usages of D-bigram Statistics in Natural Language Evaluation.
Tomohisa Sano, Junya Tsutsumi, Sun Da Jiang, Shiho Nobesawa, Kengo Sato, Kumiko Oomori and Masakazu Nakanishi.
Department of Computer Science, Keio University.

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

4 自然言語文の評価値

d-bigram を用いた相互情報量を利用して、文に対する評価値を定義する。文に対する評価値は、文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]。

$$I(W) = \sum_{d=1}^{m} \sum_{i=0}^{n-d-1} MI_d(x_i, x_{i+d}, d) \qquad (3)$$

W : 文
n : 文中の単語数
m : d-bigram の最大距離

5 d-bigram の活用方法

ここでは、d-bigram の持つ特徴を考慮しそれをどのように活用するのが適切であるかを考える。

5.1 窓かけ

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

例えば、"is" と "pen" という 2 単語 は "This is a pen ." や "This is a long pen ." のように、距離がある範囲にわたって出現すると考えることができる。

このように、ある 2 単語は距離 d の近傍に現れるというように考えるのが自然である。

そこで、式 (2) の 2 単語 x, y が距離 d で出現する確率 $P(x,y,d)$ に窓かけを行なうことで拡張する。その同時出現確率は次のように定義する。

$$P_w(x,y,d) = \sum_{w=w_{min}}^{w_{max}} \frac{C(x,y,d+w)}{M(d+w)} \times f(w) \qquad (4)$$

$$for \quad 0 \le i+d+w < n$$

w : 窓の大きさ
w_{max} : 窓の範囲の上限
w_{min} : 窓の範囲の下限
$C(x,y,d)$: 単語 x, y が距離 d で出現する回数
$M(d)$: 任意の単語が距離 d で出現する回数

$f(w)$ は 窓かけに対する重み付けであり、同時出現確率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 (図 2)。

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

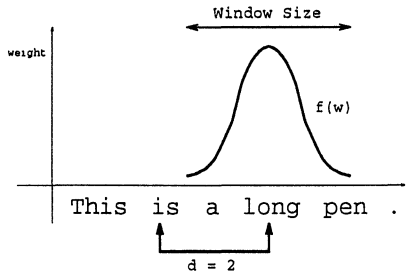

図 2: f(w) [窓かけに対する重み付け]

また、どこまでを窓かけの対象とするか、つまり、近傍の大きさについても考える。窓かけの大きさは、0, 1, 2 の 3 種類を用いる。近傍の大きさを小さくすると柔軟性の低い結果をもたらすと考えられるが、窓かけの大きさを大きくすると柔軟性が高くなる反面、曖昧性の高い結果になると思われる。

5.2 d-bigram を用いた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。つまり、距離の離れた 2 単語が偶然にある特定の距離で多く出現した場合にはその相互情報量は大きくなってしまう。しかし、距離が離れている時はその 2 単語が直接は関係のないものである (複文における前半部分と後半部分、重文における主文の部分と関係節の部分など) 可能性がある。このように、距離が遠くなるほど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が次のように考えられる。

$$I_d(W) = \sum_{d=1}^{m} \sum_{i=0}^{n-d-1} MI_d(x_i, x_{i+d}, d) \times g(d) \qquad (5)$$

$g(d)$ は d-bigram を用いた相互情報量に対する重み付けで、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 (図 3)。

$$g(d) = 1 \quad (constant)$$
$$g(d) = 1/d \quad (linear)$$
$$g(d) = 1/d^2 \quad (square)$$
$$g(d) = 1/e^d \quad (exponent)$$

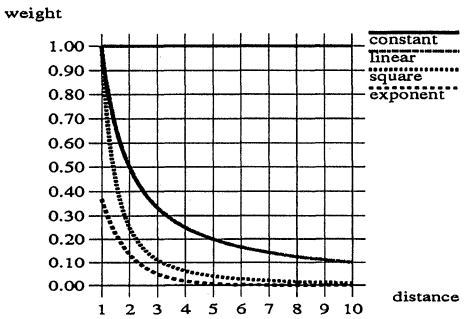

図 3: g(d) [d-bigram を用いた相互情報量に対する重み付け]

6 実験

本研究では、トレーニングコーパスとして Brown Corpus を用いた。このコーパスは、総単語数 約 120 万語、語彙数 約 4 万語、文数 約 5 万文である。

テストコーパスとして、単語数 6 以上 8 以下の文をニュースなどからランダムに選び、それぞれ 50 文ずつ計 150 文用意した。各入力文に含まれる全ての単語の全ての並べ方の組合せに対して式 (5) を用いてそれぞれの評価値を計算し、並べ方を評価値で順位付けした。窓かけの重み付け、窓かけの大きさ、相互情報量の重み付けを変えて実験を行なった。

7 結果

以下にあげた結果は、入力した文と全く同じ文が得られた場合のみを正解としている。また、窓かけの大きさが 0 の場合、窓かけをしないのと同じ意味である。

表 1 は、窓かけと相互情報量に対する重み付けに関する実験の結果である。横軸を相互情報量に対する重み付け、縦軸を窓かけの重み付けの種類とする。

表 1: 窓かけと相互情報量に対する重み付け (各 150 文)

		α	β	γ	α	β	γ
		constant			linear		
0		60	100	132	75	109	136
constant	1	52	92	129	65	104	136
	2	45	88	125	62	99	133
linear	1	57	96	130	71	105	138
	2	53	91	130	67	105	136
square	1	57	98	132	73	107	138
	2	59	98	131	73	107	137
exponent	1	57	98	132	74	107	138
	2	58	97	131	73	107	137
		square			exponent		
0		75	108	136	74	110	136
constant	1	71	101	135	67	102	135
	2	66	101	135	62	99	135
linear	1	74	106	135	72	107	136
	2	73	104	135	71	105	136
square	1	75	106	135	73	108	136
	2	75	106	135	72	108	136
exponent	1	75	106	135	73	107	137
	2	75	106	135	72	108	137

α : 正解を 1 位 に出力
β : 正解を 5 位までに出力
γ : 正解を 20 位までに出力

窓かけの大きさを 0 にした場合が良い結果を示しており、窓かけの大きさが小さいほど良い。また、相互情報量に対する重み付けでは正解が 1 位になるところでは square の場合が他よりも良い結果である。

表 2 は、窓かけの大きさを 0 にした場合の d-bigram を用いた相互情報量に関する重み付けの単語数ごとの結果である。

表 2: d-bigram を用いた相互情報量の重み付け

		6	7	8	計
constant	α	27	18	15	60
	β	40	32	28	100
	γ	50	43	39	132
linear	α	30	22	23	75
	β	39	37	33	109
	γ	50	44	42	136
square	α	31	19	25	75
	β	40	36	32	108
	γ	50	44	42	136
exponent	α	31	20	23	74
	β	40	37	33	110
	γ	50	44	42	136

α : 正解を 1 位に出力
β : 正解を 5 位までに出力
γ : 正解を 20 位までに出力

全体的に文の長さが短い (単語数が少ない) ほど正解率が良くなっていく。表 3 に 8 単語の実験の結果を示す。

表 3: 8 単語の結果 (各 50 文)

		α	β	γ	α	β	γ
		constant			linear		
0		15	28	39	23	33	42
constant	1	10	24	39	17	32	42
	2	10	23	36	16	30	40
linear	1	13	27	39	21	32	44
	2	13	23	39	20	32	42
square	1	19	29	39	21	33	44
	2	14	29	39	21	33	43
exponent	1	14	29	39	21	33	44
	2	14	29	39	21	33	43
		square			exponent		
0		25	32	42	23	33	42
constant	1	21	30	41	19	30	41
	2	19	30	41	18	30	41
linear	1	23	32	41	22	33	42
	2	23	31	41	22	31	42
square	1	24	32	41	22	33	42
	2	24	32	41	22	33	42
exponent	1	24	32	41	22	33	43
	2	24	32	41	22	33	43

α : 正解を 1 位に出力
β : 正解を 5 位までに出力
γ : 正解を 20 位までに出力

本実験では、入力文と等しい出力が得られた場合を正解としている。しかし、ある単語の集合から生成可能な "意味のある文" は入力文と同じもののみとは限らない。入力文とは異なるが意味を持つ文も正解と認めた場合、単語数 8 (50 文) 、窓かけの大きさ 0 、相互情報量の重み付け square では、表 4 のような結果が得られる。また、その実行例の一部を 図 4 に示す。

表 4: 意味のある文

	δ	ϵ
1 位	25	31
5 位まで	32	38
20 位まで	42	47

δ : 入力と同じ出力の時正解
ϵ : 意味のある文の時正解

8 評価

前節から、窓かけの大きさが 0 の時に良い結果が得られているので、窓かけは必要ないと考えることができる。これは、窓かけを行なうことによって生じるノイズの悪影響が大きいことを示している。窓かけを行なうとしても square もしくは exponent で重み付けし、近傍から外れるほど 2 単語間の関係に及ぼす影響をなるべく小さくするようにすべきである。

```
14 << @@ but go ahead and try anyway . >>
----------------------------------------
OK  (01) @@ but anyway go ahead and try .
OK  (02) @@ but anyway try and go ahead .
    (03) @@ go ahead and try but anyway .
    (04) @@ try and go ahead but anyway .
==> (05) @@ but go ahead and try anyway .
```

図 4: 実行例

d-bigram を用いた相互情報量に関する重み付けは、全体的に見ると linear, square, exponent とだいたい同じであるが、正解が 1 位にきた場合を見ると square が良い結果を出している。つまり、距離が遠い 2 単語の関係は距離が近い 2 単語に比べ、ノイズを含んでいる可能性が高いと考えられる。

今回の実験ではトレーニングコーパスとテストコーパスとして全く種類の異なるものを用いたが、それでも相互情報量の重み付けを square とし窓かけの大きさを 0 とすると、1 位 に 50 %、5 位までには 72 %、20 位までには 91 % の正解が得られた。この結果は入力と同じ出力が得られた時を正解としているが、意味のある文を集計すればさらに正解率の高い結果が得られることが予想される。

よって、d-bigram 情報を用いる際には、窓かけは行なわず相互情報量に対する重み付けを square にするのが最も効果的であると考えることができる。

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, 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.