d-bigramと他の統計情報との関連に関する実験

堤 純也, 孫 大江, 延澤 志保, 佐藤 健吾, 佐野 智久, 中西 正和
慶應義塾大学理工学研究科計算機科学専攻

平成8年 2月 28日

1 はじめに

d-bigram[1]は、単語間統計情報の一つとして種々の応用実験から [2] [3] [4] 、その有用性が示されてきた。抽出が容易であり, しかも意味的な情報を示すことができるため, コーパスの規模にあまり影響を受けない統計情報として利用価値は高い. 本論文ではbigram, trigram, d-bigram, 距離つきMI等との比較実験について報告する. 主な比較点として, 同一コーパスからの抽出情報量の差, また特にtrigramについてはd-bigramとの情報量としての差を検証することを目的とする.

2 統計情報

自然言語処理で用いられている統計情報は、以下の2種に分類される。

- マルコフ過程を基本とするモデル (n-gram系)
- 相互情報量を基本とするモデル (MI系)

n-gram系モデルは良く使用されており、中でも3単語を対象とするtrigramは良い性質を示す統計情報として様々な応用がなされてきている。また、近年では4単語以上を対象とするようなn-gramモデルを使用する研究も盛んに行なわれている [5]。n-gramモデルはcontextの情報を有効に表現できるという特徴を持つが, その反面, 距離を置いた意味的な情報には弱い. そのため, 有継文字列の切り出しや, 統語情報の抽出等には有効であるが, 係り受け等の意味的な情報抽出には使用されていない.

MI系モデルについても共起特徴を示す情報として良く利用されてきている [6]。共起関係は, <動詞-名詞> の関係など, 2単語が同時出現する確率を表現するものであるため, 通常は意味的な関係を示す.

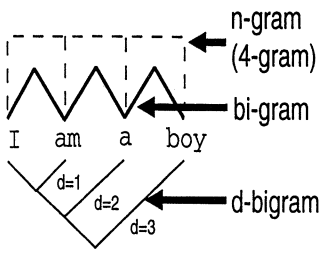

図 1: d-bigram情報とn-gram情報

今回対象とするd-bigramは基本的にはMI系のモデルであり、純粋なMI、距離ウィンドウ付きMIのどちらをも完全に包含するモデルである。また、n-gram系のモデルについても、bigramについては距離1でのd-bigramととらえられ、完全にd-bigramに包含される (図1)。bigram以上の一般的n-gramはd-bigramと完全に重なるものではないが、隣接n単語間の評価値をd-bigramの組みで計算することにより、n-gramに近い情報を与えられると考えられている。d-bigramは以下の式 (1) で表現される。

$$MI_w(x_i, x_{i+d}, d) = \sum_{w=w_{min}}^{w_{max}} \log_2 \frac{MI_d(x_i, x_{i+d+w}, d+w)}{f(w)} \quad (1)$$

x_i	:	入力列中の i 番目の要素
d	:	2 要素間の距離
$P(x)$	:	要素 x が現れる確率
$P(x,y,d)$	:	要素 x,y が距離 d で現れる確率
w	:	窓の大きさ
w_{max}	:	窓の範囲の上限
w_{min}	:	窓の範囲の下限
$f(w)$	:	窓内の重み付け関数

このように MI 系については非常に似た特徴を持つ d-bigram であるが、実際の統計情報を用いた応用の多くは以下のような特徴を持つ trigram モデルを利用することが多い。

- コーパスからの抽出が比較的容易であること
- 比較的濃密な情報が抽出可能であること

そこで、今回の実験ではこの trigram を対象とし、以下のような項目について検討を行なう.

1. コーパスから抽出される情報量の差
2. コーパスから抽出される情報の特徴

前者については、同一コーパスから trigram、d-bigram を抽出後、両者の数的優位性について検討する。まず, 第一の実験として, trigram 中の有意な上位データに対して d-bigram がどのような値の分布を示すかを検証する. 次に第二の実験として, 逆方向である, d-bigram から評価値を計算済みの 3 単語対について, ある閾値をもって trigram データととらえられるかどうかを検証する. これらの実験により, おおまかな d-bigram と trigram 間の関係をとらえることができ, 両者に相関が認められるかどうかを検証できる.

後者については, d-bigram から評価値を用いて取得可能な 3 単語対と, コーパスから取得した trigram が示す値 (頻度) がそれぞれどのような関連を示すかを検証するものである. 理想的には trigram の頻度が大きいものほど d-bigram での評価値が大きくなって欲しいわけであるが, 実際には両統計情報が得意とする情報種は異なると予想できる. 実際に本実験を行なうことで, どのような関係が認められるかを検証することは trigram, d-bigram の関係をある程度考慮できれば非常に有効であると考えられる.

3 d-bigram の特徴

実際に良く使用されている trigram モデルであっても、理想的に全 trigram 情報を取得しようと試みると莫大な記憶容量 ($O(n^3)$, n は単語数) が必要となり、またそのような全 trigram 情報を取得するためのコーパスは現実的に不可能である。抽出した結果に関しても、trigram は統語的な情報には非常に有効であるが、意味的な制約を処理することは難しい。それに対し、d-bigram は理想的に全 d-bigram 情報を取得しても記憶容量が比較的少なく ($O(n^2)$)、単語間の意味的な制約を得ることもできる [1] [3]。

上記のような理想的な状況はほぼ考えられず, 通常は与えられたコーパスからいかに有用な情報を取得できるかが問題となる. trigram(n-gram) の抽出に関しては効率の良い手法がいろいろと考案されており, 計算量に関しては比較的問題なく抽出が可能であるが, 得られる trigram の数はコーパスに顕著に依存してしまう. 次項でその数については検証する.

4 trigram の d-bigram による評価値計算

4.1 実験

本プロジェクトで使用されている、以下の代表的なコーパスについて trigram の抽出, d-bigram の抽出を行なった。

表 1: 対象コーパス一覧

コーパス名	言語	語彙数	総単語数
Brown Corpus	英語	約 50,000	約 1,200,000
PH(文字単位)	中国語	約 7,000	約 4,000,000

取得した trigram について d-bigram を用いて評価値を計算する。評価値計算については、本プロジェクトで標準的に用いている d-bigram を用いた文評価値計算、式 (2) を利用している

[1] [2] [3] [4] 。

$$I(W) = \sum_{d=1}^{d_{max}} \sum_{i=0}^{n-d-1} \frac{MI_w(x_i, x_{i+d}, d)}{g(d)} \qquad (2)$$

x_i : 入力列中の i 番目の要素
d : 2 要素間の距離
W : 入力列
n : 入力列の要素数
d_{max} : d-bigram の最大距離
$g(d)$: 距離に対する重み付け関数

4.2 結果 1

実験で trigram, d-bigram を取得した結果, 以下のような結果となった.

表 2: trigram 取得数

trigram	Brown	PH
頻度 100 以上	55	2378
頻度 10 以上	3486	50500
頻度 2 以上	75507	417130

表 3: trigram 上位 10 位の例

```
658:  ,-AND-THE
400:  ONE-OF-THE
335:  THE-UNITED-STATES
319:  ,-HOWEVER-,
266:  ,-IN-THE
252:  ,-HE-SAID
234:  AS-WELL-AS
234:  ,-IT-IS
222:  ,-AND-HE
220:  OF-COURSE-,
```

上記の trigram のうち, 頻度 100 以上のものだけについて手作業で有意なデータを抽出すると,

表 4: 有意 trigram 数

Corpus	trigram(頻度 100 以上)
Brown Corpus	20
PH Corpus	978

という結果となる. いずれのコーパスもコーパスの規模としては中規模のものであるが, 抽出データは非常に薄いものとなってしまっている. 一方, d-bigram データを用いて評価値算出をした結果は以下のようになる.

表 5: d-bigram 評価値

評価値	Brown	PH
閾値 20 以上	5890	13828
閾値 10 以上	23894	898399
閾値 0 以上	∞	∞

表 6: d-bigram 上位 10 位の例

```
15.54 THE-UNITED-STATES
12.18 HE-DID-NOT
12.17 AS-WELL-AS
11.48 HE-HAD-BEEN
11.11 THE-SAME-TIME
10.27 A-COUPLE-OF
10.17 FOR-EXAMPLE-,
10.10 IT-WOULD-BE
9.71 THERE-WAS-NO
9.45 THERE-IS-NO
```

trigram に比べて非常に大きな数値となっている. これら抽出結果についての検証は残念ながら今回の実験には間に合わなかったが, trigram と同様に上記データ中から, 手作業で有意なデータを抽出する必要がある.

4.3 結果 2

実験で取得した trigram について、出現頻度順に並べ、その順に d-bigram を用いて trigram の評価をした場合の分布が図 2、図 3である。図 4 は Brown コーパス中で出現頻度が上位 60 単語について無作為に 3 単語熟語を作成し、それを評価値計算したものである。図から分かる通り、trigram になるような結合度の高い単語列については d-bigram による評価はすべて正の値をとるという良い結果が得られ

た。一方、無作為3単語熟語については評価値が正の値をとることは少ない。

trigramとd-bigramが予想に反し秩序のない分布をしているのは、両者が取得する情報のベクトルが異なり、trigramは統語情報を抽出し、d-bigramは意味情報を抽出したためではないかと思われる。また、Brown Corpusでは評価値が低い近辺にノイズが出ているが、これはtrigramを取得する際にd-bigramと単語処理が異なっているものがあったため、未出現単語として処理されたために異常に低い評価値をとってしまったものである。

今後の検討項目として、大局的な判定だけでなく、特徴的なケース毎の両者の細かな特徴検討を行なうことで、より両者の特徴を明らかにすることが可能であると思われる。

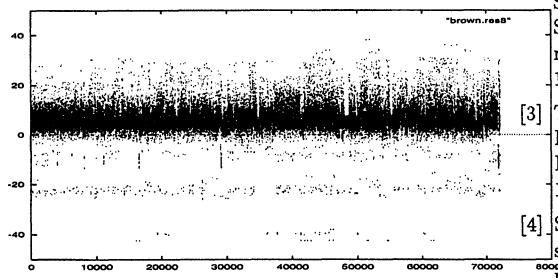

図 2: trigram 評価値 (Brown Corpus)

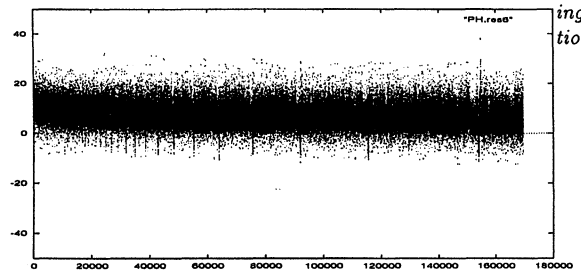

図 3: trigram 評価値 (PH Corpus)

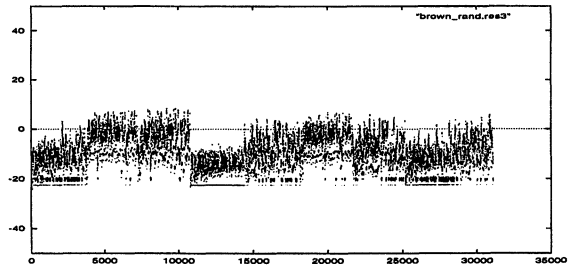

図 4: trigram 評価値 (Random3 単語列)

参考文献

[1] 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.

[2] Nobesawa, S., Tsutsumi, J., Nitta, T., Ono, K., Sun, D. J. and Nakanishi, M. Segmeting a Japanese Sentence into Morphemes Using Statistical Information between Words. *Coling*, pages 227–233, 1994.

[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.

[4] Sun, D. J., Tsutsumi, J., Nitta, T., Ono, K., Nobesawa, S. and Nakanishi, M. An intelligent Chinese input system using statistical information between words. *Qualico*, pages 102–107, 1994.

[5] 長尾 眞, 森 信介. 大規模日本語テキストの n グラム統計の作り方と語句の自動抽出. 人工知能学会 自然言語処理 *Vol.96, No.1*, pages 1–8, 1993.

[6] Church, K. and Hanks, P. Word Association Norms, Mutual Information, and Lexicography. *In Proceedings of the 27th Annual Conference of the association of Computational Linguistics*, 1989.