d-bigramとtrigramの相関に関する実験

1H－3

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

1. はじめに

本論文では単語間統計情報の一つである d-bigram[1] を、trigram と比較実験した結果について報告する。現在まで、d-bigram を用いた種々の応用実験から [2] [3] [4]、その有用性は示されてきた。本実験では d-bigram が理論上覆うことのできない [1] とされる trigram に対する d-bigram の特徴を実験的に捉えることにより、d-bigram の統計情報としての信頼性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、基本的には以下の 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系モデルは古くから良く使用されてきており、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行なわれている [5]。また、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]。

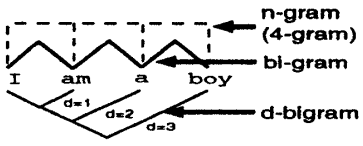

図 1: d-bigram 情報と n-gram 情報

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ものである。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)。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、今回の実験はこの trigram を対象として、d-bigram の特徴を明らかにし、そこに何らかの関連性を発見することを目的とする。

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、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取得不可能である。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]。

Experiment on Relationship between D-bigram and Trigram
Junya TSUTSUMI, Da Jiang SUN, Shiho NOBESAWA, Kengo SATO, Tomohisa SANO, Masakazu NAKANISHI
Department of Computer Science, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa Pref., 223, Japan

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

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。

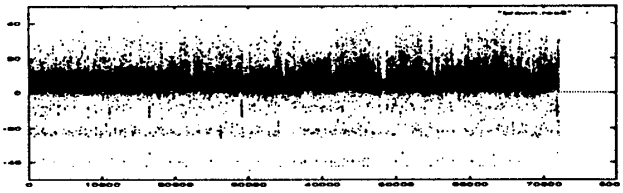

図 2: trigram 評価値 (Brown Corpus)

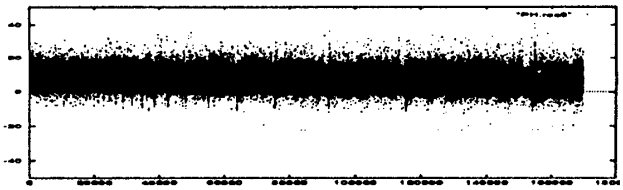

図 3: trigram 評価値 (PH Corpus)

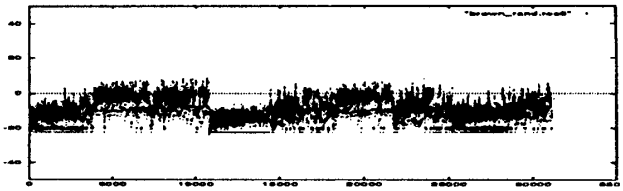

図 4: trigram 評価値 (Random3 単語列)

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.