大容量テキストのn－gram統計とその応用の検討[1]

4E－4

豊浦潤[2]
RWCP[3]新機能三菱研究室[4]

1　はじめに

様々なソースから流入するリアルタイム情報の自然言語処理方法が問題になっている。リアルタイム情報は、情報サイクルが短く、語彙が経時的に増大するため、従来主流であった辞書やルールベースを利用した自然言語処理モデルでは、対応が困難である。そこで近年、統計情報を利用した自然言語処理モデルが注目されている。

自然言語処理に対し統計情報を利用するモデルは、従来からも音声認識などでは成功を収めていたが、日本語テキストでは使用される文字種が多いため、適用は困難視されていた。しかし、近年の計算機の進歩により、日本語に関しても、統計的に十分大きな量のテキストから統計情報を求めることが、可能になってきている[1]。

だが、現時点では、統計情報の利用に際し「どの統計量が有効なのか」「標本とするテキストの量はどの程度必要なのか」に対する検討は十分行なわれてはいない。そこで、筆者は、数万件の新聞記事に対しn-gram統計をとり、上の問題を考察したので以下に報告する。

2　テキスト統計情報

テキストに対する統計情報としては様々なものが考えられるが、語彙の増加など言語のダイナミクスに対応するという意味では、情報の単位を小さくとった統計量が有効である。そこで、テキストに対し、n-gram 統計をとることにした。

今、長さ n の文字列を、$X_n = x_1, x_2, \cdots, x_n$、また、$L$:$X$ を構成するアルファベット数、$P(X_n)$: 長さ n の文字列中の X_n の出現確率、と定義する。

このとき n-gram 統計に対し、1 文字当たりの結合エントロピー:G_n、1 文字当たりの条件付きエントロピー:F_n は、以下のように定義される。

$$G_0 = F_0 = \log L$$

$$G_{n\geq 1} = \frac{1}{n} \sum_{x_1,\cdots,x_n} P(X_n) \log \frac{1}{P(X_n)}$$

$$F_{n\geq 1} = \frac{1}{n} \sum_{x_1,\cdots,x_n} P(X_n) \log \frac{1}{P(x_n \mid X_{n-1})}$$

情報が定常であれば、G_n, F_n は n の増大に従い単調減少し、$\lim_{n\to\inf} G_n = \lim_{n\to\inf} F_n = H$ が存在することが知られている[2]。

ところが想定している新聞などは、定常な情報ではなく、L や n-gram 数は、時間の経過に伴い増大する。しかし、十分多数の X について、n-gram をとれば、時間変化に対する揺れが小さな漸近値が得られることが、期待できる。

一方、L には 上限値:L_{max} が存在する。よって、n-gram 数の上限は理論上は、$(L_{max})^n$ となるが、これは、各文字が独立である場合であり、現実のテキストでは n-gram 数の漸近値は理論値よりずっと小さいことが予想され、十分大きな n についての n-gram は漸近値を持つことが期待される。これらの問題を解決するために、以下の実験を行なった。

3　実験

3.1　準備

実験用のテキストデータとして、朝日新聞の記事を用いた（表1）。

掲載年月日	1990年度1月～8月
総記事数	62555
総文字数	(598.7/記事)
異なり文字数(L)	4498

表 1: 実験用テキストデータの諸属性

n-gram 抽出は、SS10 上で、jperl のプログラムを実行した。ディスク容量の関係で、今回は $n = 1 \sim 5$ まで抽出することにした。

3.2　X の増加に対する収束性

1月1日から8月31日までの記事を順に蓄積したときの $G_0, \cdots, G_5$ 変化を図1に示す。G_n は、X に関

[1] A study on n-gram statistics of large text and its application
[2] Jun TOYOURA
[3] Real World Computing Partnership
[4] Novel Functions Mitsubishi Laboratory

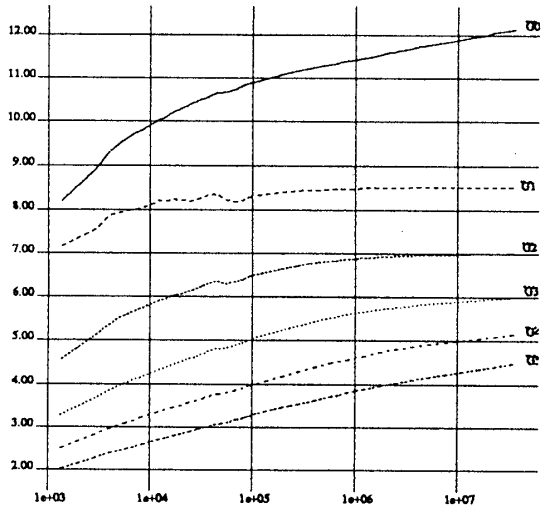

図 1: テキスト量増加に伴う G_n の変化

し単調増加であるが、その増加率は指数関数より早く減少しているが、調査した範囲では G_n が収束に向かうのかは不明である。しかし、$G_1, \cdots, G_2$ に関しては 1月中に、$G_0, G_3, \cdots, G_5$ に関しても 4月中には、最終値（8月終了時）の 99% に到達しており、現実的にはほぼ収束しているとみなせる。

3.3 n の増大に対する収束性

n の増大に対する F_n, G_n をプロットしたグラフを図2に示す。日本語は英語と比べ、F_0 は、かなり大き

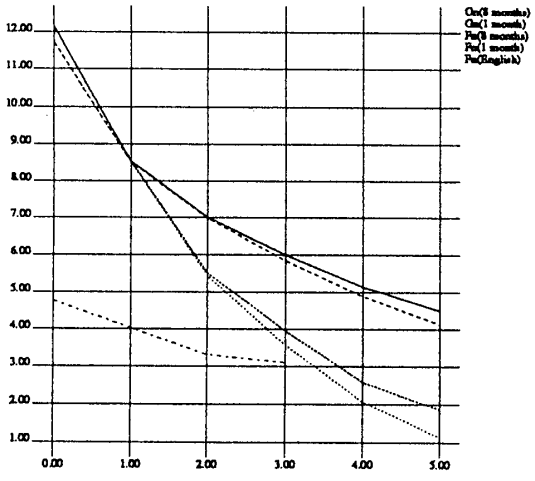

図 2: n 増大に伴う F_n, G_n の変化

いが、今回求めた $n = 0 \sim 5$ では急激に減少し、$n = 4$ では、ほぼ等しくなる。つまり日本語は、小さい n-gram で、新たに 1 文字を知った時の獲得情報量は大きいが、大きい n-gram での獲得情報量は小さい。

今回調査では F_n は収束していないが、上限値は $H < 2$ が分かった。日本語の H も、英語の推定値:$H = 1.2 \sim 1.6$ に近い値をとると仮定すれば、F_n は $n = 7, 8$ 近辺でほぼ収束すると予想できる。

4 応用

ここまでは大局的に日本語の統計的性質を見てきたが、本節では、その応用を考える。マルコフ確率は形態素解析などで、文字間の接続率を評価する際に用いられることが多いが、異なる文字のマルコフ確率を単純に比較するだけ接続率の評価にならない。そこで、前方・後方の条件付きエントロピーの比で評価することを考える。

図3は幾つかの文字について、bi-gram から求めた、前方条件付きエントロピー、後方条件付きエントロピーでプロットしたものである。図の右下には、「前の文字

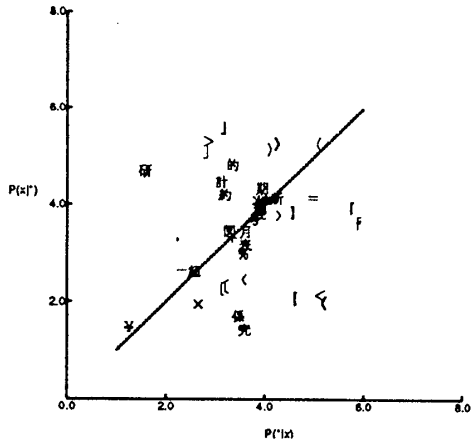

図 3: 1 文字に対する前方・後方条件付きエントロピー

は予想できるが、後の文字は分からない」文字が、左上には、「前の文字は分からないが、後の文字は予想できる」文字が集まっていることが分かる。

5 まとめ

大量のテキストの n-gram の統計的性質を考察し、その応用を検討した。今後は、さらに応用範囲を拡張する予定である。

参考文献

[1] 長尾 他, 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, 情処N L研資, *96*, 1, pp.1-8, 1993.

[2] 例えば、佐藤洋, 情報理論, 基礎物理学選書, 15, 裳華房, 1973.