2007－NL－182 (7)
2007／11／19

日本語文章における順方向と逆方向の文字遷移情報量の差異

中川正樹，朱碧蘭
東京農工大学

本稿では，自然言語における文字遷移の情報量として順方向と逆方向を考え，それぞれについて前後2文字の文字遷移の情報量を遷移確率上位の有限個だけ加算した場合，日本語では順方向と逆方向で顕著な差異があることを示す．また，加算する有限個の遷移確率値の総和が１になるように正規化した場合の正規化情報量でも同様の傾向があることを示す．

Difference in the amount of information between forward and backward transitions in Japanese text

MASAKI NAKAGAWA and BILAN ZHU
Tokyo University of Agriculture and Technology

This paper defines the amount of information for forward transition and that for backward transition for natural languages and measures them in terms of 2-gram. Then, it presents that a notable difference exists between them in Japanese text when the amount of information is accumulated from the most probable transition to a limited number with descending transition probabilities. The same difference is also observed between the forward transition and the backward transition in normalized accumulated information when the sum of accumulated transition probabilities is normalized to 1.

1. はじめに

自然言語は，形態素列として，あるいは，文字列として見ることができる．本稿では後者の立場から，日本語文章を文字列として捉え，文字遷移の確率（n-gram確率），特に，前後２文字の文字遷移確率 （2-gram確率）をもとに，文字遷移の持つ情報量を求め，日本語の文字遷移に関する新たな知見を提示する．

自然言語の情報量については，情報理論に対する最初期の論文[1]から多く扱われてきている．日本語に対しても，文字列に対して効率的にn-gram統計を得る手法の提案[2]，文字認識後処理のために文字連接の有無の可能性を見る初期の研究[3]，形態素列の確率モデル[4][5][6]，文字列の確率モデル[7][8]，それらを融合したモデル[9]が提案されてきた．また．音声認識やタイプ入力なども含めて日本語文章の誤り検出と訂正手法の提案もある[10]．包括的なサーベイは西野[11]，基礎理論は北[12]に詳しい．基本に立ち返って，形態素単位の n-gramモデルによる日本語情報量の上限を推定する研究[13]，文字列のn-gramモデルを用いた著者推定の研究[14]もある．

現在では多様な日本語コーパスが利用可能になり，むしろこうした分野の発表は少なくなり，一方で，実利用が進んでいる．分野自体はこうした状況にあるが，本稿では，文字認識のための文脈後処理を検討する過程で明らかになった新たな知見を報告する．それは，日本語における文字遷移の情報量として順方向と逆方向を考え，それぞれについて前後2文字の文字遷移の情

報量を遷移確率上位の有限個だけ加算した場合，順方向と逆方向で顕著な差異があることを示すものである．荒木らは，2重マルコフモデル（3-gram）を用い，順方向（前の二つから最後），逆方向（後ろの2つから最初），中間（最初と最後から中間）の誤り音節／文字訂正能力を比べている[15]．しかし，ここで提案するように，候補を有限個に限定した場合の情報量は検討していない．

2. 文字遷移の情報量

2.1 順方向と逆方向の情報量

文字列中における任意の2-gramの1文字目をS_i，2文字目をS_jとして，S_iからS_jへの遷移が持つ情報量は，単位をbitとして，次式で表される．

$$I(S_j \mid S_i) = -\log_2 P(S_j \mid S_i) \qquad (1)$$

ここで，文字遷移の情報量として，順方向と逆方向の２種類を考える．特に，2-gramについて考える（図１）．2-gram S_iS_jが出現する確率を$P(S_iS_j)$として，文字列を順方向に走査したときにS_iからS_jへの遷移が持つ平均情報量（エントロピー）をH_F，逆方向に走査したときにS_jからS_iへの遷移が持つ平均情報量をH_Bとすると，H_F，H_Bは

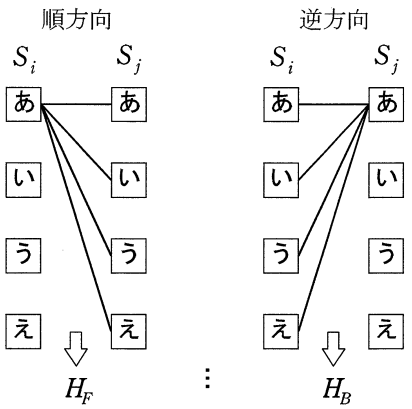

図１．情報量計算の２つの方法

それぞれ次のように表される．

$$H_F = \sum_{\forall i, \forall j} \left(P(S_iS_j) \cdot -\log_2 P(S_j \mid S_i) \right) \qquad (2)$$

$$H_B = \sum_{\forall i, \forall j} \left(P(S_iS_j) \cdot -\log_2 P(S_i \mid S_j) \right) \qquad (3)$$

以下，情報量の単位はすべてbitとする．

2.2 累加情報量

式(2)と(3)を，すべての文字遷移の組合せについて計算すれば，まったく同じ計算をすることになる．しかし，ここでは，情報量の加算を，それぞれ$P(S_j|S_i)$または$P(S_i|S_j)$の高い方からm位までに限定する．これを加算限定m位までの累加情報量（Accumulated Information）と呼ぶことにする．つまり：

$$\tilde{I}_F(m) = \sum_{\forall i, P(S_j|S_i)\text{の上位 } m\text{個の } j\text{について}} \left(P(S_iS_j) \cdot -\log_2 P(S_j \mid S_i) \right) \qquad (4)$$

$$\tilde{I}_B(m) = \sum_{\forall j, P(S_i|S_j)\text{の上位 } m\text{個の } i\text{について}} \left(P(S_iS_j) \cdot -\log_2 P(S_i \mid S_j) \right) \qquad (5)$$

上位有限個までの累加情報量を考える理由は，現実の文字認識後処理などへの文字遷移確率の利用において，ほとんどの場合，文字認識における有限個の上位候補に対してしか文字遷移確率を考えないことによる．

累加情報量は，加算限定mが文字種数に近づくにつれて平均情報量に漸近していく．

2.3 正規化累加情報量

式(5)，(6)において，$P(S_j|S_i)$または$P(S_i|S_j)$を上位m個しか加算しないことから，加算限定mが文字種数より小さい間はこれら確率値の総和は1にはならない．そこで，加算対象の確率の総和が1になるように正規化して，累加情報量を検討する．こうすれば，その大小で，確率分布がど

の程度均等か，あるいは，不均等かを判断できる．

そこで，次のように，$P'_m(S_j \mid S_i)$，$P'_m(S_i \mid S_j)$を定義する．

$$P'_m(S_j \mid S_i) = \frac{P(S_j \mid S_i)}{\sum_{P(S_j|S_i)\text{の上位}m\text{個の}j\text{について}} P(S_j \mid S_i)} \quad (6)$$

$$P'_m(S_i \mid S_j) = \frac{P(S_i \mid S_j)}{\sum_{P(S_i|S_j)\text{の上位}m\text{個の}i\text{について}} P(S_i \mid S_j)} \quad (7)$$

これらを用いて，順方向および逆方向の正規化累加情報量（Normalized Accumulated Information）をそれぞれ$\tilde{H}_F$，$\tilde{H}_B$として，次のように定義する．

$$\tilde{H}_F(m) = \sum_{\forall i, P(S_j|S_i)\text{の上位}m\text{個の}j\text{について}} \left(P(S_i) P'_m(S_j \mid S_i) \cdot -\log_2 P'_m(S_j \mid S_i) \right) \quad (8)$$

$$\tilde{H}_B(m) = \sum_{\forall j, P(S_i|S_j)\text{の上位}m\text{個の}i\text{について}} \left(P(S_j) P'_m(S_i \mid S_j) \cdot -\log_2 P'_m(S_i \mid S_j) \right) \quad (9)$$

3. 実験と考察

3.1 実験対象

日本語の情報量の計算には，朝日新聞記事データベース（CD-HIASK'93）から抽出したテキストを用いて作成した2-gram統計表を用いた．このテキストは約110MBあり，その先頭から1/1000（約110kB），1/100（約1MB），1/10（約11MB）を抜粋したものを，それぞれテキストとして統計表を作成した．なお，日本語文字1文字は2バイトである．

日本語の情報量の特徴と比較するために中国語と英語の情報量についても調査を行った．中国語の情報量の計算には，人民日報新聞記事データベース（2002年1年分）から抽出したテキストを用いて作成した2-gram統計表を用いた．このテキストは約55MBである．英語の情報の計算には，ACM学会論文誌のTransactions on Information Systemsの1995年から2004年まで10年分の論文のデータベースから抽出したテキストを用いて作成した2-gram統計表を用いた．このテキストは約10MBである．

これらのテキストついて，前章で述べた順方向（Forward），逆方向（Backward）の累加情報量，及び，正規化累加情報量が加算限定mの増加（先頭を除いて5刻みのm=1,5,10,15,...）によって，どのように変化するかを調べた．

3.2 実験結果と考察

日本語のそれぞれのテキストについて，テキスト中に出現した文字カテゴリ数を表１に示す．

表１．テキスト中の文字カテゴリ数

テキストのサイズ	カテゴリ数
110 KB	1740
1 MB	2626
11 MB	3632
110 MB	4799

日本語のテキストサイズが110KB, 1MB, 11MB, 110MBのときの，加算限定mによる累加情報量の変化を図２に示す．横軸の最大は文字種数となるが，収束したところで打ち切る．テキストが大きくなるほど差が顕著になることから，偶然の揺らぎとは考えられない．以下のすべての統計はこの傾向を示すことから，最大テキストに対する統計だけを示すことにする．改めて，横軸の範囲を小さくし，立軸を拡大して，110MBのテキストに対する統計を図３に，そして，正規化累加情報量の統計を図４に示す．

また，加算限定mによる累加情報量と正規化累加情報量の変化を，中国語については図５，図６，英語については図７，図８に示す．横軸の最大は文字種数となるが，収束したところで打ち切る．

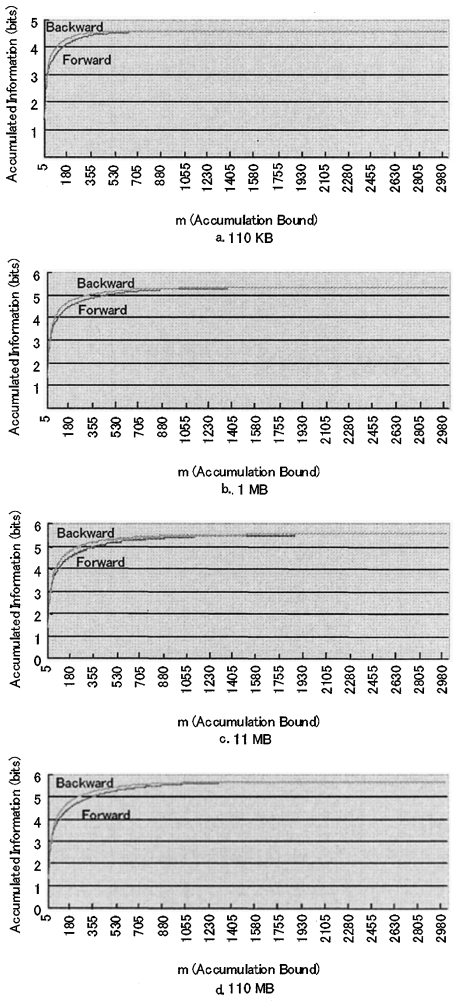

図２．日本語の加算限定に対する累加情報量の変化

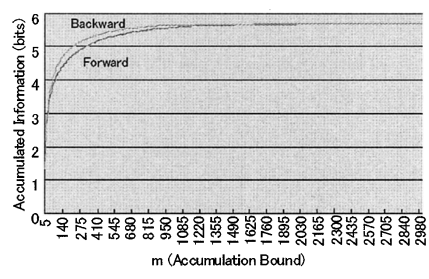

図３．日本語の加算限定に対する累加情報量の変化

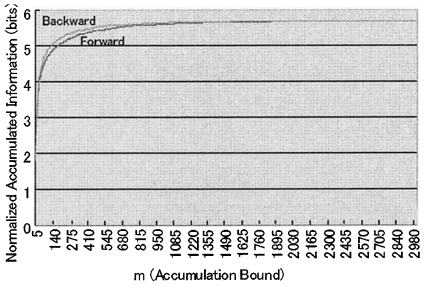

図４．日本語の加算限定に対する
正規化累加情報量の変化

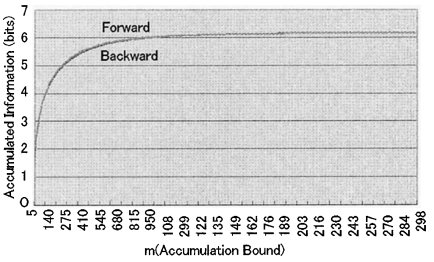

図５．中国語の加算限定に対する累加情報量の変化

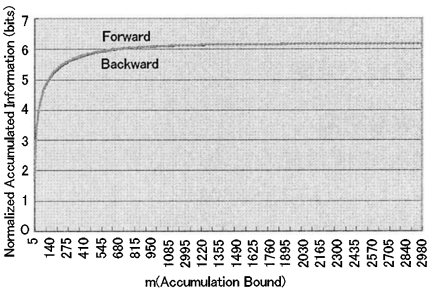

図６．中国語の加算限定に対する正規化累加情報量の変化

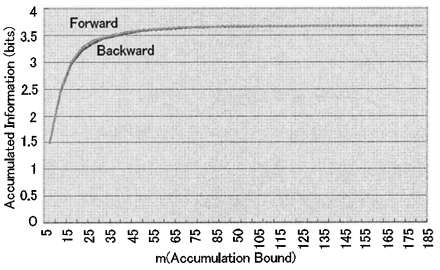

図７．英語の加算限定に対する累加情報量の変化

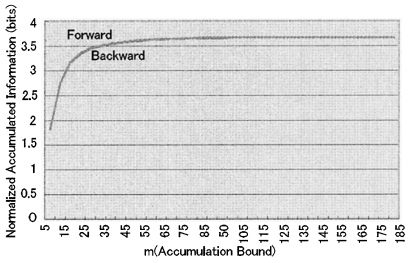

図８．英語の加算限定に対する正規化累加情報量の変化

図３，図４から，日本語のテキストにおいて加算限定mが小さいときには，順方向と逆方向で情報量に顕著な差異を示し，累加情報量，正規化累加情報量ともに逆方向のほうが順方向より大きい情報量を持つことが分かる．しかし，図５～図８から，中国語と英語のテキストでは，加算限定mと関係なく，順方向と逆方向で情報量に顕著な差異を認めなかった．

3.3 考察

加算限定mのある区域において，累加情報量，正規化累加情報量ともに逆方向のほうが順方向より大きい情報量を持つということは，逆方向のほうが順方向より偏りの少ない分布をしていること，言い換えれば，順方向のほうが逆方向より偏りが大きいことを意味する．したがって，逆向きに前の文字を予測するより，順方向に次の文字を予測するほうが容易だということが示唆される．

荒木らは，3-gramによる誤り訂正法で，順方向が誤り訂正能力で優位であることを示している [15]．その理由として，上記の言語的特性が考えられる．

我々の現状のオンライン手書き文字認識では，認識候補の中から文字遷移確率の高いパスを選択する方法であるため，順方向と逆方向で差がでない方式になっているが，逆に候補文字列から次の文字や前の文字を予測して，その認識スコアを評価する方式では，順方向が逆方向より優位になることが予想される．この検証には，認識システムの再構成が必要なため，今後の課題とする．

いずれにしても，加算限定mが大きくなるにつれて等しい値に収束していく二つの情報量が，途中の段階において顕著な差を示すことは，言語の特性を示していると捉えることができる．中国語と英語では，このような特性を認めなかった．この特性を3-gramでも調査したい．加算限定mの中途段階において，文字遷移情報量がこのような特性を示す理由の本質的な考察については，今後の課題としたい．

4. おわりに

本稿では，自然言語における文字遷移の情報量として順方向と逆方向を考え，それぞれについて前後2文字の文字遷移の情報量を遷移確率上位の有限個だけ加算した場合，日本語では逆方向のほうが順方向より大きい情報量を持つことを示した．加算する有限個の遷移確率値の総和が１になるように正規化した場合の正規化情報量でも同様の傾向があることが分かった．以上を新聞1年分の記事から確認したが，さらに別の日本語コーパスで確認すること，3-gramでも調査すること，この特性を利用できる文字認識方式と組み合わせて評価すること，本質的な理由を考察することを今後の課題とする．

謝辞　本研究は，科学研究費補助金・萌芽研究16650014の補助による．

参考文献

1) Shannon, C. E.: Prediction and Entropy of Printed English, *Bell System Tech. J.*, Vol. 30, No. 1, pp.50-64 (1951).
2) 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，情報研報，Vol. 93, No. 61, pp.1-8 (1993.7).
3) 杉村利明，斎藤珠喜：文字連接情報を用いた読み取り不能文字の判定処理－文字認識

への応用ー, 電子通信学会論文誌, Vol. J68-D, No. 1, pp.64-71 (1985).
4) 高尾哲康, 西野文人：日本語文書リーダ後処理の実現と評価, 情報処理学会論文誌, Vol. 30, No. 11, pp.1394-1401 (1989).
5) 伊藤伸泰, 丸山宏：OCR入力された日本語文の誤り検出と自動訂正, 情報処理学会論文誌, Vol. 33, No. 5, pp.664-670 (1992.5).
6) Masaaki Nagata: Context-Based Spelling Correction for Japanese OCR, Proc. 16th COLING, pp.806-811 (1996).
7) 森大毅, 阿曽弘具, 牧野正三：2重マルコフモデルを用いた日本語文書認識後処理, 情報研報, Vol. 94, No. 63, pp.89-96 (1994. 7).
8) 荒木哲郎, 池原悟, 塚原信幸：マルコフモデルを用いたOCRからの誤り文字列の訂正効果, 情報研報, Vol. 94, No. 63, pp.97-104 (1994.7).
9) 森大毅, 阿曽弘具, 牧野正三：文字・単語n-gramの融合に基づく言語モデル, 情報研報, Vol. 96, No. 65, pp.109-114 (1996. 7).
10) 荒木哲郎, 池原悟, 塚原信幸：2重マルコフモデルによる日本語文の誤り検出並びに訂正法, 情報研報 Vol. 93, No. 79, pp.29-35 (1993.9).
11) 西野文人：文字認識における自然言語処理, 情報処理, Vol. 34, No.10, pp.1274-1280 (1993).
12) 北研二：確率的言語モデル, p.239, 東京大学出版会, 東京 (1999).
13) 森信介, 山地治：日本語の情報量の上限の推定, 情報処理学会論文誌, Vol. 38, No. 11, pp.2191-2199 (1997.11).
14) 松浦司, 金田康正：n-gram分布を用いた近代日本語小説文の著者推定, 情報研報, Vol. 99, No. 95, pp.31-38 (1999. 11).
15) 荒木哲郎, 池原悟, 土橋潤也, 堂元一頼：2重マルコフモデルの全域法と局所法による日本語の誤字訂正効果, 情報研報, Vol. 93, No. 61, pp.9-16 (1993.7).