コーパス分割を用いた良質な統計量の推定

3R－7

乾 伸雄，小谷 善行
東京農工大学工学部情報コミュニケーション工学科

1. はじめに

統計的アプローチに基づく自然言語処理においては，質の高い統計量をコーパスから得ることが解析率を向上させる鍵となる．一般にコーパスの規模が拡大すると，そのコーパスに対する解析精度は低下する．これを防ぐために，「**コーパス分割法**」という新しい手法を導入し，日本語の形態素解析において評価する．

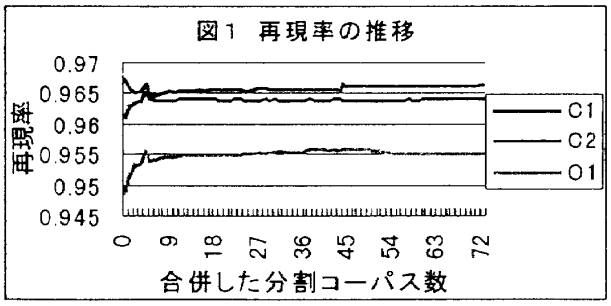
2. 学習コーパスの解析率の低下

本稿では，統計量を取得するためのコーパスを「学習データ」，それ以外のコーパスを「評価データ」と呼ぶ．形態素解析には式1を用い，学習データとしてRWCコーパス91年度（約2700万単語，品詞492種類）を使う．図1は，学習データサイズと解析率の関係を示す．学習コーパスを128分割し，おのおのを分割コーパスと呼ぶ．横軸は，合併された分割コーパス（累積分割コーパス）の数を示す．式(1)は生起確率と文脈確率の積で表されている．C1（39300 語）とC2(44781 語)は生起確率の算出に使われているデータであり，O1(65409 語)は使われていない．また，C1は文脈確率の算出に使われているが，C2，O1 は使われていない．

(式1) $\underset{i}{\operatorname{argmax}}\, P(W_i \mid P_i)\, P(P_{i+1} \mid P_i)$

W_i：i番目の単語・品詞，P_i：i番目の品詞

図1 再現率の推移



データの規模が大きくなるにつれて，C1 の再現率が低下していき，一定の値（約 210 万語のコーパス，10セットの分割コーパス，に対して，再現率 0.964）に収束した．C2，O1 については，ある時点まで再現率が向上するが，やはり収束する．収束した時点で，累積分割コーパスから得られる文脈確率が，一定値になったと考えられがちであるが，そうではない．これについては，3章で議論する．また，C1,C2とO1との差は，未知語が原因となって発生したものと考えられる．

累積分割コーパス規模が小さいとき，C1の解析精度が下がる理由としては，多品詞語処理の問題を指摘できる．規模が小さいときは，C1 に含まれる多品詞語をうまく処理できても，大きくなるにつれて，多品詞語を一意的に解釈するような結果しかだせなくなる．実際に，形態素境界の決定率の低下を再現率の低下は大きく上回っている．この意味で，文脈確率は微妙なバランスの上に成り立っているといえる．コーパス規模拡大とともに解析精度が下がる問題を防ぐために，本稿ではコーパス分割法を4章において提案し，5章で評価する．

3. 精度の推定

文脈確率が C1 に似ているほど，C1 に対する再現率は向上すると考えられる．これから，二つの文脈確率のうちで，C1 の文脈確率に似ている方が，再現率がより高いことが予想される．このため，文脈確率をベクトルと考え（式2），二つのベクトル間の非類似度を式3のような尺度で測定した．表1は再現率に対するC1 の属する分割コーパスと累積分割コーパスの非類似度の相関係数を示す．

(式2) $PLnm = (P(pos[1] \mid pos[m]), P(pos[2] \mid pos[m]), \cdots)$

$PLn = (P(pos[1] \mid pos[1]), \cdots, P(pos[1] \mid pos[2]), \cdots)$

PLnm：品詞ごとに，文脈確率をベクトル化

PLn：文脈確率全体をひとつにベクトル化

(式3) $Da1,2(1,n) = \sum_m angle(PL1m, PLnm) / m$

$|PLXm|=0$ のときは 90 度，無視

$Da3(1,n) = angle(PL1, PLn)$

$Dd1,2(1,n) = \sum_m |PL1m - PLnm| / m$

$|PLXm|=0$ のときも計算，無視

$Dd3,4(1,n) = |PL1 - PLn|$

$|PLXm|=0$ のときも計算，無視

angle(PL1m,PLnm)：コーパス L1,Ln における品詞mの文脈確率のなす角度

表1から，Da2,Da3,Dd2 の相関が高いので，頻度0の品詞については無視した方がよいことがわかる．本

Estimating Excellent Statistics by a Corpus Separation Method
Nobuo Inui, Yosiyuki Kotani
Tokyo University of Agriculture and Technology

稿では，この結果に従い，二つの文脈確率の非類似度をDa2によって，推測する．

表1 文脈確率と再現率の相関係数

Da1	Da2	Da3	Dd1	Dd2	Dd3	Dd4
-0.702	-0.79	-0.792	-0.738	-0.78	-0.748	-0.779

しかし，分割コーパス間の角度については，式(4)が近似的に成り立ち，累積分割コーパスの値は，一定の値に収束しないで振動する．全ての分割コーパスに対し，最も類似した文脈確率を求めることが望まれるが，これを一意的に定めることはできない．このため，本稿では4章で述べる手法を用い，全てのコーパスに似ている文脈確率を求める．

(式 4) $\forall i,j\ Da2(i,j) \fallingdotseq 18$

4. コーパス分割法

2，3章で述べてきたように，精度を向上するためには，どの分割コーパスにも似ている文脈確率をコーパスから推定することが必要である．従来の方法では文脈確率は式5によって計算されるが，個々のコーパスの違いをより反映するために，式6によって文脈確率を計算する．基本的には同じ文脈確率を持つコーパスを重視する(しない)という方針で計算する．freq(pos[i])は品詞 pos[i]の頻度を示し，Px は分割コーパス x についての文脈確率を表す．式5は式6の特殊な場合，つまり全てのコーパスについて品詞の頻度が等しい場合に相当する．

(式 5) $P(pos[i] \mid pos[j]) = \dfrac{freq(pos[j]pos[i])}{freq(pos[j])}$

(式 6-1) $Pz(pos[i] \mid pos[j]) = \dfrac{Px(pos[i] \mid pos[j]) + Py(pos[i] \mid pos[j])}{2}$

(式 6-2) $P(pos[i] \mid pos[j]) = \dfrac{\Sigma Px(pos[i] \mid pos[j])}{k}$

k： 残りの分割コーパスの数

指定された残りの分割コーパス数になるまで，式 6-1によって，二つの分割コーパスの文脈確率を統合し，式6-2によって，平均を求める．式6-2で統合される分割ペアの選択方法として，今回の実験では次の二つの方法を使う．

方法α)最も類似しない文脈確率を選択

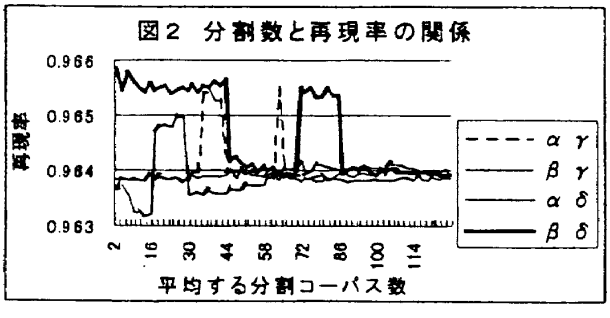
方法β)最も類似した分割コーパスを選択

非類似度は次の二つで計算できる．

方法γ)分割コーパスを単位として文脈確率を平均

方法δ)品詞を単位として文脈確率を平均

図2 分割数と再現率の関係



5. コーパス分割法の評価

図2に L1 に対する再現率を示す．横軸は，式 6-2を適用する残りの分割コーパス数である．例えば1のときは，式 6-1 だけで文脈確率を求めることになる．

β δ，つまり，非類似度が最小の組み合わせを品詞単位，累積分割コーパス数が1になるまで行なったとき，最大の再現率0.6959を得た．図1と比較すると，通常の手法による再現率の低下を100%としたとき，約40%の低下に押さえている．α γにおいて，累積分割数 60 付近で，再現率が瞬間的に向上しているが，分割コーパスの組み合わせによって様々な C1 に近い文脈確率を表現できることを示している．しかし，他の文脈確率と比べて，C1に近い二つの文脈確率の非類似度 Da2 は決して小さくない．そのため，より高精度な非類似度の尺度が必要となる．表2に，C1を最大にする文脈確率のO1に対する再現率を示した．評価データに対しても，再現率は向上している．正解率も，再現率と同程度向上することが確認できた．

表2 評価データO1に対する再現率

	通常	α γ	β γ	α δ	β δ
再現率	0.9553	0.9569	0.9553	0.9553	0.9568

6. 関連研究

精度を向上するためには，トライグラム，共起関係[1]，品詞の階層化・低信頼統計量の推定[2]などの手法が考えられている．コーパス分割法はこれらと矛盾なく利用することができ，形態素解析だけはなく，様々な統計的手法への適用が可能である．

7. おわりに

コーパス分割によって，学習コーパスの規模拡大による解析精度の低下を防ぐ方法を提案し，日本語形態素解析における効果を確認した．

謝辞 本研究は文部省科学研究費補助金(09780315)の支援で行われた．

参考文献

[1] 北研二他，音声情報処理，森北出版，1996

[2] 藤本他，枝分かれ構造を持つ同時確率モデルによる形態素解析，情報処理学会論文誌(近刊)