単語ラティスからの最尤単語列選択方式

2J-6

小黒 雅巳，中村 修，北村 正
NTT 情報通信処理研究所

1 まえがき

文字認識結果等の不確かな入力文字列から尤度の高い単語を抽出する技術は，ユーザフレンドリな入力処理を実現する上で重要である。このため，先に，漢字1単語に相当する入力文字列から精度良く単語を抽出できる連想統合型照合方式を提案した[1,2]。

本報告では，複合語に相当する入力に対しても有効な単語列選択方式を提案する。誤認識等のあいまい性を許容した単語列選択の実現には，入力文字列中の各文字から連想される単語候補の組み合せから，最尤候補を選択する必要があり多くの演算量を要する。そこで本方式では，単語列の尤度として，単語が持つスコアの他に，隣接関係および文法的接続関係を評価尺度とし，効率の良い評価法を実現する。

2 隣接関係に基づく単語列選択

本選択処理においては，入力文字列中の各文字から連想される単語候補を対象に，これらの隣接関係を検証すると同時に，各単語候補に付されたスコアの和が最高となる組み合せを選択する。上記の処理は，最大2^{m-1}（m; 入力文字列長）の組み合せについて実行する必要がある。本稿で示す方式は，以下の方法により演算量の削減を可能としている。

(a) 入力文字列中における単語候補の開始位置及び終端位置によって番地付けした配列に，対応する単語候補のスコアを配置する。

(b) 上記配列に対し，機械的なアドレス歩進によって，隣接関係の検証ならびにスコアの加算の対象とすべき単語候補の組み合せを得る。

(c) 上記(b)のスコア加算の結果から，各文字位置までの低スコア候補をふるい落とす。

本方式の処理例として，複合語"情報処理研究"の誤認識により"情結処理研究"が入力され，連想統合処理[1]の結果図1(a)に示す単語ラティスが得られた場合の処理を図2に示す。具体的な処理手順は以下となる。

(1) 配列要素$[i,j]$に開始位置i，終端位置jの単語候補$W[i,j]$およびそのスコア$S[i,j]$を配置する。

(2) 配列要素$[i,j]$において，配列要素$[i-1,i-1]$が出力するスコアと$S[i,j]$との加算結果と，配列要素$[i-1,j]$が出力するスコアとのうち高い方を選択する。図2中の配列要素[4,4]の例では，"情報処理"が選択される。

上記(2)の処理を，i,j共に2からmまで繰り返す。

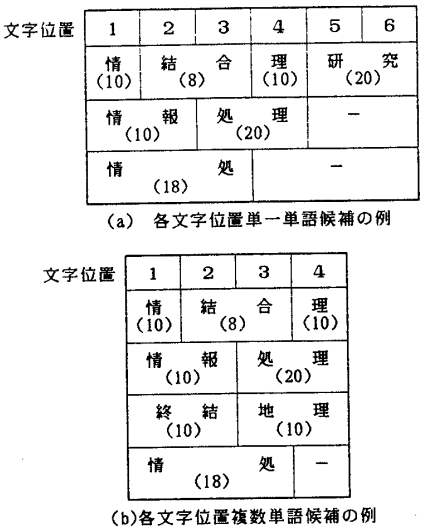

図1 単語ラティスの例

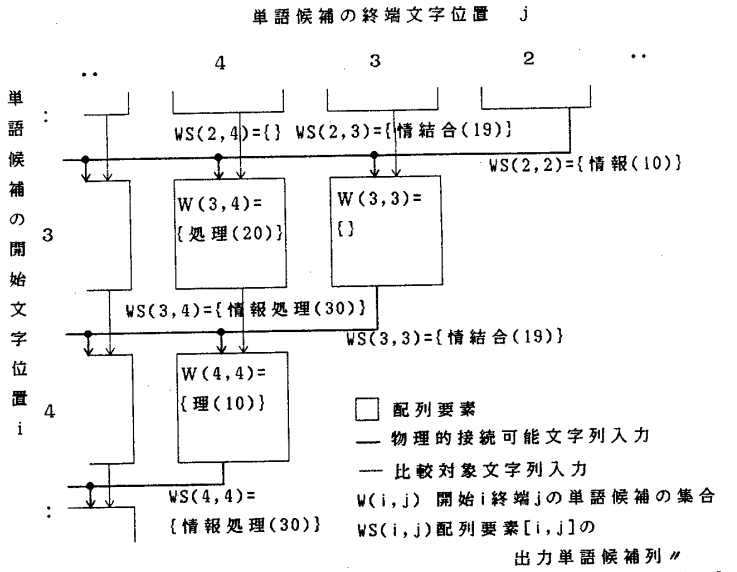

図2 物理的接続関係に基づく選択法

A best word selection method for a word lattice
Masami Oguro, Osamu Nakamura, Tadashi Kitamura
NTT Communications and Information Processing Laboratories

3 文法的接続関係適用への機能拡張

隣接する単語候補に対して，文法的な接続関係が規定できる時，各文字位置において，同スコアの部分単語列や，図1(b)に示す複数の単語候補に対する絞り込みを上記の文法的接続関係を用い実現できる。具体的には，以下に示す方法によって，2章で示した機能を拡張する。

(a) 2章(c)に示した単語候補のふるい落としに際して，上記の文法的接続関係の検証を加える。

(b) 各配列要素内の単語候補の属性より，あらかじめこれらの単語候補に関連する部分接続規則を求め，隣接関係にある単語列が入力されると単語列の属性で部分接続規則を参照し接続関係を検証する。

図3は，図1(b)に示す単語ラティスを対象に，上記の文法的接続関係の検証を含めた単語列選択処理を示している。具体的な処理手順は以下の通りである。

(1) 配列要素$[i,j]$に配置された単語候補$W[i,j]$に付された属性で，接続規則を参照し，関連する部分接続規則のテーブルを作る。図3の例では，属性B,Cについてのテーブルが求められる。

(2) 配列要素$[i-1,i-1]$からの出力単語列の属性でテーブルを行方向に参照し配列要素$[i,j]$の単語候補との文法的接続関係を検証する。図3では，属性AとCが接続可能なため，"情報処理"が選択される。

(3) 上記(2)までで得られた単語列のスコアと配列要素$[i-1,j]$が出力するスコアとのうち，高い方を選択する。

上記(1)～(3)の処理を，i,j共に2～mまで繰り返す。

4 演算量削減効果

隣接関係に基づく単語列選択，および，文法的接続関係を適用した場合の単語列選択方式について，演算量の削減効果を求めた。効果の算出に当たっては，抽出された全単語候補について接続し得る全ての組み合せのスコアを求め，その中から最尤候補を選択する方法（全数法）との比較を行った。

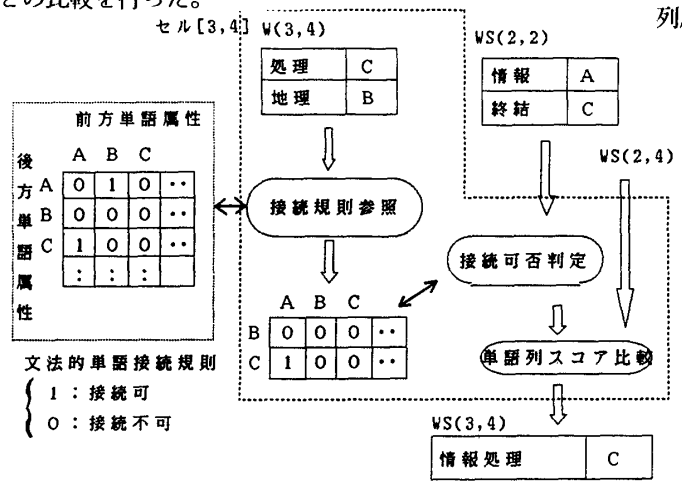

図3 文法的接続関係に基づく選択法

隣接関係に基づく単語列選択方式の効果算出条件

全数法P_a = 隣接単語候補組数N_c × 加算時間T_a + 全単語候補列数N_s × 比較時間T_c

$$= T_a \cdot \sum_{i=1}^{m} 2^{i-1} + T_c \cdot \sum_{j=1}^{m-1} (^{m-1}_{j})$$

本方式P_p = 配列要素数$N_e \times (T_a + T_c)$

$$= (T_a + T_c) \cdot \sum_{i=1}^{m-1} i$$

文法的接続関係適用時の効果算出条件

全数法$G_a = N_c$×検証すべき単語数N×(T_a+照合時間T_m)+ $N_s \times T_c$

$$= N_{i,j} \cdot N_{i-1,i-1} \cdot (T_a + T_m) \cdot \sum_{i=0}^{m} 2^{i-1} + T_c \cdot \sum_{j=1}^{m-1} (^{m-1}_{j})$$

本方式$G_p = N_e \times (T_a + T_c$+接続可否判定時間$T_j) + N_e \times$ 接続規則参照時間T_f

$$= N_{i-1,i-1} \cdot (T_j + T_a + T_c) \cdot \sum_{i=1}^{m-1} i + N_{i,j} \cdot T_f$$

$N_{i,j}$：配列要素$[i,j]$の単語候補数

算出結果

図4に，上記の算出条件から求めた本方式の演算量削減効果を示す。図4から以下の結論が得られる。

(1) 全数法より約1桁の高速化が期待できる。

(2) 上記の効果は入力文字列数の増加に伴い大きくなる。

5 むすび

単語候補の隣接関係および文法的接続関係を評価尺度として，複合語相当のあいまい文字列から効率良く単語列の選択が可能な方式を提案した。本方式は，複合語に対する連想統合型照合方式として活用できる他，単語間接続関係を文法的に規定できる場合[3]には，さらに精度の良い単語列選択が可能である。また，本方式は，加算，大小比較等の基本機能の繰り返しにより実現可能で，並列展開，専用ロジック化が容易という利点も有する。

文献

[1] 松尾他：第34回情処全大予稿4E-7(1987).

[2] 仲林他：第1回人工知能全国大会予稿8-7(1987).

[3] 宮崎他：情処論文誌 Vol.25 No.6(1984).

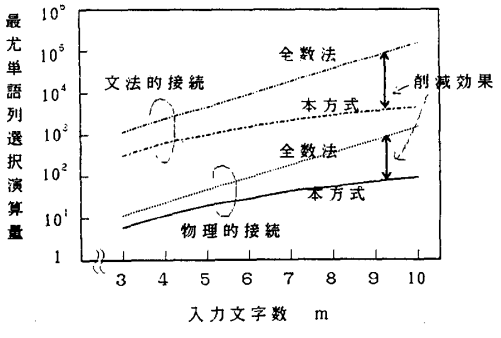

図4 演算量削減効果