画像処理手法による文字列分解型の文字列検出方法の提案

4F−3

津村　和弘，小林　大助，池田　旬
（株）東芝

1．まえがき

単語・熟語、及びこれらを組合わた造語を用いた効率の良い（自然な書式）知識表現で知識を構築し、自然語文字列（記号）例えば、「あるぜんちんはどこ」の入力に対して、これに含まれる前記知識の構成文字列「あるぜんちん」「どこ」を容易（形態素解析等によらない）に検出できるならば、知識の作成、処理の高度化に有効である。図書及び内容検索を目的として、文字列処理に画像処理手法を適用した検索システムを構築し有効性を確認した。

2．「表」による知識表現

人が利用している「表」は、簡潔で判りやすい知識の表現形態で、アルゼンチンの地理に関しても、「表」に準じて、以下（表1）のように表現できる。このような固有名詞とか、「りんご」「腐敗」いった物・現象，行動等を直接表現した文字列は、人が最も多く使用する文字列である。このような文字列は、物・現象，行動等をより具体化した現実との対応データ（下位レベル）の他に、「南米独立国」とか、「果物」といった上位概念の文字列に関連して知識を構成している。

南米独立国	大陸・地域	所在地	所在地質問$	アルゼンチン
アルゼンチン ＝あるぜんちん ブラジル	南米，北米 アフリカ アジア	地域 場所	どこ	緯度／−22〜−55 気候／ 隣接国／ブラジル

表1　「表」に準じた表文書の一例

上位概念の文字列（知識）は、知識を効率良く整理したり、同一概念の元に整理された知識（例えば内容別等に分類整理された図書、及び（目次から）関連ページを検索する場合に必要である。つまり、任意の自然語文字列を構成する意味ある文字列（単語・熟語等）を判断し、質問に答えたり、関連した図書・ページを検索し提示するためには、単語・熟語、及びこれらを組合わた造語または文を効率良く整理した知識に対して、効果的な記号処理技術を開発する必要がある。

このためのキーポイントは以下の2点と考えた。

① 単語・熟語・造語等で表現された知識文書の文字列で、自然語文字列を意味ある部分文字列に分解し、

② 部分文字列に対応する上位概念の文字列を取出すことである。一方、目的とする文書全体に対して文字列検出が実施できる状態では、より具体的に表現された下位レベルの文字列を取出すことである。

本システムでは、単語・熟語、及びこれらを組合わた造語を効率良く整理した知識表現を、表形態（文書）として構築した。

3．画像処理手法による文字列分解型の文字列検出方法

本方法は複数の画像メモリに、前記整理された表文書（常駐知識と呼ぶ）と、

Development of Knowledge-based Algorithm for Words Extraction from a Japanese Phrase using Image Processing Technique.

Kazuhiro TSUMURA, Daisuke KOBAYASHI, Jun IKEDA
TOSHIBA, Corp.

必要に応じて利用される一般図書（非・常駐図書と呼ぶ）をコード化文書として格納しておき、データ変換，積和演算，論理フィルタ等の画像処理機能を用いて自然語文字列を構成する単語・熟語等に対応させた数値データとして、別の画像メモリに生成するものである。画像メモリに格納されているコード化文書のＣＲＴ表示（輝度表示と文書のダンプ表示）の一例を図１に、処理の概念図を図２に示す。ここで、16bit 長の日本語文字のデータを１画素８bit の画像メモリに書込む方法として、Ｙ方向に固定値（0），続いて文字コードの上位８bit、下位８bit を連続して格納する方法とした。以下、図２の処理について説明する。

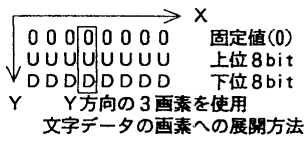

図1　コード化文書のＣＲＴ表示

① まず最初に、入力文字列を構成する各々文字コード（16 bit）の上位，下位データをデータ変換によって、より小さいデータ（1，2，3…）に変換（最小化）する。

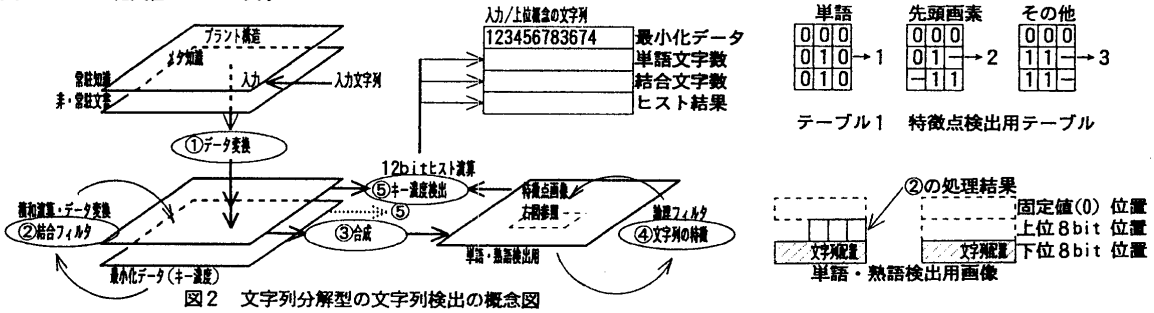

図2　文字列分解型の文字列検出の概念図

② 積和演算・データ変換による結合フィルタで、入力文字列の文字データをもとに、上位と下位（または隣合った）データが正しく結合しているデータを、再度最小化（最初は1，2，3 …をキー濃度と呼ぶ）して取出す。

③ 候補文字を示す②の処理結果と常駐知識のコード化文書から単語・熟語検出用の特徴点画像の初期画像を合成する。

④ ③の画像に対して、テーブル１の論理フィルタで、１画素（１文字）及び候補文字の先頭画素（文字）とその他を識別した特徴点画像を作成する。

⑤ ④の特徴点画像を上位４bit、②の出力を下位８bit とするヒスト演算で、単語・熟語及び結合文字のキー濃度を検出する。このヒスト結果から、図２に示す単語及び結合文字数等の情報を設定する。

次に、文字列を構成する連続した任意の２文字の結合を検出する処理として、②④⑤の処理を行う。このとき②で付与するキー濃度は、前回の処理で単語・熟語として検出された場合等を除いて、前回と同じキー濃度とする。

以上の処理は、ヒスト結果の度数に変化がなくなった場合、終了とする。

4．まとめ

本方法は、自然語文字列の文字数に関係なく（並列的）文字列に含まれる単語・熟語等が検出可能で、かつ全体処理を基本とした単純なアルゴリズムであるため、並列処理，パイプライン処理が容易である。今後、画像処理手法による記号処理の適用範囲を拡大し、システムの機能を充実させたい。