3Q－6

元文書のレイアウト情報に基づく文書構造解析手法

平沼義直　畑山佳紀　竹山哲夫
三洋電機株式会社 ハイパーメディア研究所

1. はじめに

パソコンの普及やインターネットの発展に伴い、電子メールに代表される電子化文書が急速に増加している。このような状況下では、膨大な文書から必要な情報を効率よく取得したり、正確な文書をより早く作成し発信することが要求され、その実現には、文書の持つ論理構造を自動抽出する文書構造解析技術が必要となる。

本稿では、元文書のレイアウト情報を利用し、従来手法で解析できなかった文字列の意味を推定する文書構造解析手法について述べる。

2. 文書構造解析

2.1 従来手法

従来の文書構造解析手法では、1)形態素解析により元文書を単語レベルに分割し意味を付加する、2)意味の並びから文、段落レベルの意味を導き出し「タイトル」「見出し」といった文書の構成要素を決定する、3)構成要素間の主従関係を決定する、といった手順で文書の論理構造を抽出していた。

しかし、「～について」「～のご案内」のような「タイトル」のキーワードとなる単語を含んでいない場合は「タイトル」であると推定することは困難であった。また、「作成者」「差出人」などは固有名詞が多く含まれるため手順 1)での意味付けが正しく行われず、解析誤りが多かった。

一方、元文書を作成する側の視点に立つと、スペース、タブコード、あるいは改行コードを適宜挿入して、「タイトル」はセンタリング、「作成者」は右揃えといった具合にある程度体裁を整えながら作成していくことが多い。

従って、元文書の行揃え、字下げ桁数などのレイアウト情報には論理構造の抽出に有用な情報が含まれていると考えることができる。

2.2 レイアウト情報を取り入れた解析手法

まず、元文書から各行の行揃え属性を抽出する。第 n 行の文字開始桁位置を S_n、文字終了桁位置を E_n、各行の桁数のうち最も大きなものを元文書の幅 W としたとき、図 1 のルールにより行揃え属性を決定する。

条件	行揃え属性
$S_n > W/2$	右揃え
$E_n - S_n \leqq W/2$ AND $E_n \geqq W \times 0.8$	
$S_n \geqq W \times 0.4$ AND $E_n \geqq W \times 0.9$	
$W \times 0.45 \leqq (S_n + E_n)/2 \leqq W \times 0.55$ AND $S_n \geqq W \times 0.1$ AND $E_n \leqq W \times 0.9$	センタリング
上記以外	左揃え

図 1 行揃え属性の決定ルール

従来手法の手順 2)の後、各行に付加された意味と行揃え属性から新しい意味を推定する。図 2 に新しい意味の推定ルールの一部を示す。

A Method of Analyzing Document Structure Using Original Layout Information
Yoshinao Hiranuma, Yoshinori Hatayama, Tetsuo Takeyama
Sanyo Electric Co.,Ltd. Hypermedia Research Center

行揃え属性	手順 2)の結果	新しい意味
センタリング	行の一部に「文書番号」を含む 行の一部に「宛先」を含む 行の一部に「差出人」を含む 行の一部に「担当」を含む 行の一部に「タイトル」を含む	文書番号 宛先 差出人 担当 タイトル
	前行が、「文書番号、日時、宛先、差出人、作成者、文頭」のいずれか	タイトル
右揃え	行の一部に「宛先」を含む 行の一部に「差出人」を含む 行の一部に「作成者」を含む 行の一部に「担当」を含む 行の一部に「タイトル」を含む	宛先 差出人 作成者 担当 タイトル
	最後もしくは最後から2番目の行であり、「氏名、部署、電話番号」のいずれかが含まれている	担当
	前行が、「文書番号、日時、宛先、表記、差出人、作成者、文頭」のいずれかであり、「氏名、部署、組織」のいずれかを含む	作成者
	先頭もしくは先頭から2番目の行であり、「数字列」を含む	文書番号
	前行が「宛先」	差出人
左揃え	先頭から4番目以内の行であり、行の一部に「宛先」を含む	宛先

図 2 行揃え属性による意味推定ルール

3. 本手法の評価

無作為に抽出した社内文書 10 通を対象として、従来手法と本手法の解析精度の評価を行なった。その結果を表 1 に示す。

手順 1)および手順 2)の処理で用いる形態素辞書(約 56,000 語)、意味構造辞書(約 2,000 ルール)は両手法とも同じものを使用した。

表中の「抽出構成要素」は文書構造解析により抽出した文書の構成要素数、「正解要素」は「抽出構成要素」のうち正しい意味が付加された構成要素数を計数したものである。そして「解析正答率」を「正解要素」／「抽出構成要素」により算出した。

本手法により、従来手法で解析誤りとなっていた構成要素のうち約 39％が正解となった。また、本手法により、逆に解析誤りとなってしまった構成要素はなかった。

表中の「抽出構成要素」に差異が生じたのは、従来手法において正しく解析されなかった構成要素が「本文」と認識され、続く行の「本文」要素に取り込まれてしまったからである。

表 1 解析結果（対象文書：社内文書 10 通）

従来手法	抽出構成要素	145
	正解要素	121
	解析正答率	83.4%
新手法	抽出構成要素	147
	正解要素	132
	解析正答率	89.8%

4. おわりに

元文書のレイアウト情報を利用することで従来手法で解析できなかった文字列を推定する文書構造解析手法を提案した。従来手法では解析誤りの多かった「タイトル」や「作成者」などの構成要素を正しく解析できることを確認した。

今後は、さらに解析精度を向上するためレイアウト情報による意味推定ルールの最適化を行なっていく。