文書構造解析における知識ベースの自動構築

3J−2

矢島尚子　坂東博司　黄瀬浩一　日下浩次
大阪府立大学

1.はじめに

文書構造解析とは，文書画像を領域分割し，得られた部分領域に，題目，著者名などの論理オブジェクト名をラベル付けする処理である．現在，我々のシステムでは，様々な文書に対する適用性を保持するため対象文書に依存した部分を知識ベースに分離，蓄積する形式をとっている．ところが，現システムでは，知識ベースの生成を人手に頼っているため，かなりの労力がかってしまう．そこで、知識ベースの自動構築が必要となってくる．本稿では，与えられた文書例から知識ベースを自動構築する方法について述べる[1]．本手法の特徴は知識ベースをインクリメンタルに更新する点にある．

2.知識ベース

ここで扱う知識ベースは論理構造，レイアウト構造，対応規則の3つからなる．

2.1論理構造

論理構造とは，論理オブジェクトからなる木構造のことをいう．本手法では，論理構造が一定のものについて知識ベースを生成するため，論理構造は与えられているとする．

2.2レイアウト構造

レイアウト構造とは，文書画像を構成する文字列領域やブロック領域 (レイアウトオブジェクト)の包含関係を表す木構造のことをいう．本手法では，木構造を明示的に表現するためフレームを用いて知識を記述する．各フレームはレイアウトオブジェクトに対応する．以後，これをレイアウトオブジェクトフレーム (LAOF)と呼ぶ．LAOFは，部分全体関係と類似差異関係により結合されている．部分全体関係はレイアウト構造における包含関係 (親子関係)を規定するものである．また，類似差異関係は子領域間の差異を規定するものである．類似差異関係については，LAOFとは別に類似差異フレーム(SIMF)を設けて記述する．

表1,2にLAOFとSIMFの各スロットと内容を示す．ここで，layout_descriptor スロットに記述される特徴量記述子とは，特徴量名と区間値の組により表されるものである．LAOFの場合は，横幅や縦幅など各レイアウトオブジェクトの特徴量を記述する．また，SIMFの場合は，距離など各レイアウトオブジェクト間の類似差異を表す特徴量を記述する．また，レイアウト述語はflush(寄せ)，alignment(揃え)などのレイアウトを表す記号である．

表1 LAOFのスロット

スロット	内容
name	各レイアウトオブジェクト固有の名称
type	レイアウトオブジェクトの種類
number_of	レイアウトオブジェクトの個数
order	親が共通のレイアウトオブジェクトの並ぶ順序
part_of	親のレイアウトオブジェクトの名称
subpart	子のレイアウトオブジェクトの名称のリスト
orientaiton	子のレイアウトオブジェクトの並ぶ方向
similarity	SIMFの名称のリスト
layout_descriptor	特徴量記述子，レイアウト述語のリスト

表2 SIMFのスロット

スロット	内容
name	SIMFの名称
layout_descriptor	特徴量記述子，レイアウト述語のリスト

2.3対応規則

レイアウト構造から論理構造への対応付けは，レイアウトオブジェクト名を左辺に，論理オブジェクト名を右辺に記述したルールにより表す．

3.知識ベースの自動構築

本手法は，文書画像の例を入力とし，フレーム形式の記述 (インスタンス記述)に変換後，適当なクラスに分類し，クラスを特徴付ける記述 (クラス記述)を求めるものである．ここで，インスタンス記述とは，文書画像の例に含まれるレイアウトオブジェクトをLAOF, SIMFにより記述したものである．また，クラス記述とは，知識ベースに相当するものである．以下では，インスタンス記述とクラス記述の生成方法について述べる．

3.1インスタンス記述の生成

レイアウトオブジェクトは，領域を表す座標，レイアウトオブジェクトの種類，論理オブジェクト名で表現されている．まず，これらのデータから各レイアウトオブジェクトに対応するLAOFを生成する．ここでは，個々のレイアウトオブジェクトのデータから求められるスロットのみ記述し，残りは，空欄にしておく．次に，レイアウトオブジェクトの包含関係から各LAOFにに部分全体関係のリンクを張る．最後に，同一の親をもつフレーム間にSIMFを設けて類似差異関係を記述する．

Automatic Construction of Knowledge Base for Document Image Analysis
Naoko YAJIMA　Hiroshi BANDO　Koichi KISE　Hiroji KUSAKA
University of Osaka Prefecture

3.2クラス記述の生成

まず，構造マッチングと呼ぶ処理により，インスタンス記述とクラス記述をマッチングし，インスタンス記述の属するべきクラスを決定する．次に，記述の一般化と呼ぶ処理により，構造マッチングの結果に基づいて，クラス記述を生成，修正する．属するべきクラス記述がないか，照合するクラス記述がない場合は，インスタンス記述を一般化し，新しいクラス記述とする．また，属するべきクラス記述がある場合は，インスタンス記述を特徴付けるようにクラス記述を一般化する．以下に，構造マッチング，記述の一般化について詳細に述べる．

(1)構造マッチング

LAOF の照合をレイアウト構造の根から順に親子関係に沿って再帰的に行なう．照合済みの親 LAOFをp,qとし，pの子LAOFをx_i，qの子LAOFを y_jとする．但し，p,x_iをインスタンス記述のLAOF，q,y_jをクラス記述のLAOFとする．これらのLAOFに対し，(i) p,qの子の並ぶ方向が一致するか，(ii) x_i,y_jの出現順序の小さい順にレイアウトオブジェクトの種類と論理オブジェクトが一致するかを照合する．(ii)では，図1に示すように，y_j のレイアウトオブジェクトの種類がブロック領域であれば，個数が一致しなければならない．また，文字列領域の場合は，1個の y_j に対し，連続する限り複数の x_iと照合する．照合の可能であったものは，(x_1,y_1)のようなマッチペアとして保存する．マッチペアが得られなければ照合を中断し，次のクラス記述と照合をする．

(2)記述の一般化

数え上げ，条件削除，区間拡張と呼ぶ3種類の規則を用いて，フレームを一般化する．数え上げは，インスタンス記述の一般化に用いられる．連続して並ぶ複数の文字列領域に対し，LAOFを1つにまとめる．このとき，number_ofスロットはまとめたLAOF の個数の合計に変更する．条件削除は，一般化する2つのフレーム (F1, F2)に記述されたレイアウト述語を対象とし，F1, F2に共通するレイアウト述語を，一般化して得られるフレーム(F3)のレイアウト述語とする．区間拡張は，F1, F2に記述された特徴量記述子を対象とし，各特徴量の区間値を包含する区間値をF3の特徴量記述子に記述する．

新たにクラス記述を生成する場合は，同一のインスタンス記述のフレームに対し，数え上げ可能なら，数え上げ，条件削除，区間拡張をする．また，マッチペアが得られた場合は，その各々に対して，次のように一般化をする．クラス記述のLAOFとインスタンス記述のLAOFとが1対1の対応のときは，条件削除と区間拡張を行なう．1つのクラス記述のLAOFが複数のインスタンス記述の LAOFと照合できた場合は，インスタンス記述に対して数え上げを行ない，これとクラス記述の LAOFを条件削除，区間拡張により一般化する．

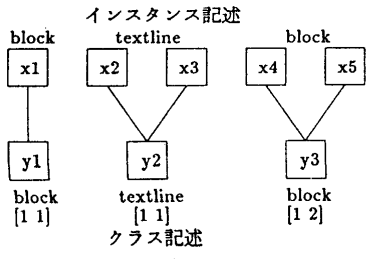

図1 照合可能なペア

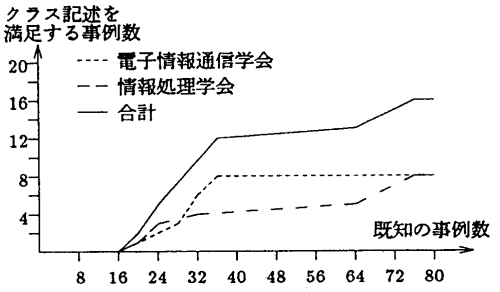

図2 実験の結果

4実験と検討

実験は電子情報通信学会論文誌，情報処理学会論文誌のタイトルページ各50 枚，計100 枚に対して行なった．このうち各40サンプルを交互に呈示することにより，クラス記述を逐次更新し，その度，残りの20 サンプル (未知サンプル) のうちいくつがクラス記述を満たすかを調べた．結果を図2に示す．クラス記述の更新に用いるサンプルの数が増加するに従って，クラス記述を満足する未知サンプルの数が増加している．このことより，クラス記述の一般性がインクリメンタルに向上することがわかる．

また，最終的に得られたクラス記述を知識ベースとし，文書構造解析を行なった．評価の尺度として信頼度と平均仮説数を用いた．平均仮説数は，レイアウトオブジェクト1個あたりに生成されるレイアウトオブジェクトの数である．信頼度は，生成されたレイアウトオブジェクトに正解が含まれる割合である．実験の結果，平均仮説数が1.0，信頼度は99.6%となった．このことから，本手法で得られた知識ベースは，文書構造解析において有効であるといえる．

5.おわりに

本稿では，文書構造解析システムにおける知識ベースの自動構築の手法について述べた．本手法は，逐次呈示される文書画像の例をインクリメンタルに分類し，クラス記述を更新，生成することを特徴としている．また，実験ではサンプル数の増加に伴なうクラス記述の一般性の向上と，得られた知識ベースの有効性を示した．

参考文献

[1] 黄瀬他：文書構造解析のためのレイアウト構造のクラスタリング，画像の認識・理解シンポジウム論文集I, pp.287-294 (1992).