自然言語処理 115－20
情報学基礎 43－12
（1996．9．13）

SDAIRに見る文書解析と情報検索に関する研究状況

日立製作所 中央研究所
藤澤浩道

要旨：ネバダ大学ラスベガス校の情報科学研究所 (UNLV-ISRI) が主催する「文書解析と情報検索に関するシンポジウム」(SDAIR: Symposium on Document Analysis and Information Retrieval) に見る，これら分野の研究状況について報告する．'96年で第５回になった本会議は，OCRを用いたイメージ文書の検索技術を中心的課題とし，文書認識と情報検索の二つの分野の研究者が会する数少ない貴重な会議である．回を重ねる毎に，これら二つの分野の技術の有効な融合が見られる．
キーワード：印刷文書，文書理解，文字認識，OCR，情報検索，文書分類

Researches on Document Analysis and Information Retrieval Seen in SDAIR

Central Research Laboratory, Hitachi, Ltd.
Hiromichi Fujisawa

Abstract: Reviewed here are researches presented at a series of SDAIR's, "Symposium on Document Analysis and Information Retrieval," which have been sponsored by Information Science Research Institute of University of Negada, Las Vegas. The series of SDAIRs, focusing on OCR applications and information retrieval techniques especially for large-scale document image bases, have been an invaluable forum for the researchers from these two technical fields. Year by year, a more effective merge of the two technologies has been happening.
Keywords: machine-printed document, document analysis, character recognition, OCR, information retrieval, document classification, information extraction

１．はじめに

近年，益々進む文書の電子化に伴い，印刷文書の重要性はむしろ増してきている．特に，印刷文書の蓄積と検索のためのシステム，ひいてはそのための文書認識と情報検索技術の必要性が増している．

これに呼応して，ネバダ大学ラスベガス校 (UNLV) の情報科学研究所 (ISRI) は，1992年に「文書解析と情報検索に関するシンポジウム」(SDAIR: Symposium on Document Analysis and Information Retrieval) の主催を始めた．OCRを用いたイメージ文書の検索技術を中心的課題とし，文書認識と情報検索の二つの分野の研究者が一堂に会する数少ない貴重な会

議である．論文件数は毎年約２５件に絞り，質の高い論文を集めている．

ちなみに，本会議のタイトルでは「文書解析」(document analysis)となっているが，日本では，文書認識 (document recognition) ないしは文書理解 (document understanding) ということが多い．本報告では，用語「文書認識」を使う．

毎年，ラスベガス市内の会議場に150～200名の参加者が，３日間の日程で，これら二つの分野から集まる．今年で第５回になった本会議は，回を重ねる毎に，これら二つの分野の技術を融合した新しい技術が見られるようになっている．

本会議の一つの特徴は，UNLV-ISRIが毎年行うOCRの公開コンテストの報告である．毎年６社から８社が参加して，約1500頁（約300万文字）の印刷文書の認識実験を行い，各種の詳細な評価を行っている．

本報告では，上記OCRの公開コンテストを含め，本会議における５年間の発表内容をレビューし，この分野の研究状況を概観する．

２．発表論文の概況

発表論文は以下の４つに大きく分類できる．各年の発表内容の分類を表１に示す．

A) 文書認識：文字認識，レイアウト解析，単語認識などを含み，特に情報検索に係わらない文書認識固有の技術．手書文字認識，帳票認識などはここに分類する．

B) 情報検索：文書分類，概念検索，対話型検索，ユーザインタフェースなどを含み，イメージ文書を特には対象としない検索固有の技術．

C) イメージ文書検索：OCR結果を用いた情報検索，OCRを用いないイメージ文書の検索手法，イメージ文書の自動分類など．検索を意図した低品質文書イメージの認識技術もここに含める．

D) その他：図面認識，楽譜認識，顔画像検索，言語の同定，など．

表１　発表論文の分類

	分類	92年	93年	94年	95年	96年
A	文書認識	10	6	10	6	8
B	情報検索	9	9	7	8	6
C	イメージ文書検索	4	7	5	9	7
D	その他	1	5	3	2	1
	合計	24	27	25	25	22

印刷文書をスキャンしてイメージとして蓄積・検索するシステムにおける常套手段は，イメージをOCRで認識，コード情報 (ASCIIデータ) に変換して，それを蓄積・検索する方法である．しかし，OCRによる認識は完全ではあり得ない．従って，認識の不完全性の問題を軽減するための認識後処理方式，検索方式，或はそのうまい組合せが研究対象である．

すなわち，上記の分類で，C) が本会議で最も中心的な課題である．この分類の研究論文が増えていることは，この会議が初期の目的を達成しつつあることを示している．本報告では，主にこの分類の研究について述べる．

3．OCRコンテスト

実は，ネバダ大ラスベガス校の情報科学研究所(UNLV-ISRI)は，米国エネルギー省 (DOE: U. S. Department of Energy) の基金を元に1990年に設立された．その目的は，以下の3つである [1].

1）印刷文書理解の応用研究の遂行

2）SDAIRの主催

3）既存OCR技術の評価

この第3の目的のためにUNLV-ISRIでは，膨大な英語文書イメージのテストデータベースを構築し，公開のOCRコンテストを93年から毎年行っている．テストデータも毎年増えており，第1回コンテストでは，2500文書（10万頁）からランダムに選んだ500頁を対象にしたが，第3回では，1500頁を対象にしている．評価の方法は，参加社（参加者）がワークステーションで走るソフトウェアの形でOCRをISRIに提供し，ISRIが実験評価するという形態を取る．

これまで，Caere (93, 94, 95), Calera Recognition Systems (93, 94), Xerox Imaging Systems (93, 94, 95), ExperVision (93, 94), HP Laboratories (95), Recognita America (93, 94), Electronic Document Technology (Syngapore) (94, 95) など，合計13社が参加している（カッコ内数字は参加年）．

テスト文書は300 dpiでスキャンし，画像品質に従って，ほぼ同じ量の5つのランクの集合 (Group 1-5) に分類している．評価は以下の項目を含み，かなり詳細かつシステマティックである [1, 2, 3].

- 文字認識率 (93, 94, 95)
- 単語認識率 (93, 94, 95)
- 画像品質と精度の関係 (93, 94, 95)
- 単語長と精度の関係 (93, 94, 95)
- リジェクトフラグ (marked character) の有効性 (93, 94, 95)
- スキューの影響 (94)
- 解像度の影響 (94)
- 文書の種類による違い (95)
- ストップワード／非ストップワードの精度比較 (95)
- スペイン語文書の認識精度 (95)
- 認識速度と精度の関係 (95)
- 中間調画像を用いるOCRの認識精度 (95)

評価の結果，以下の事実が明らかになっている．

文字認識率は，高品質画像の文書 (Group 1) に対しては，どの商用OCRも99.3%以上の精度を出した．上位3種のOCR (ExperVision, Calera, Caere) は，Group 1に対して，それぞれ99.90%, 99.92%, 99.88%，Group 2に対して，99.71%, 99.56%, 99.63%であった [1]．全体の誤読文字の内の約70%は，品質の悪い20%の文書で発生している．また，OCRの違いは，むしろ，低画質の文書 (Group 5) に対してであり，これら上位のOCRは認識率が93%にしか落ちないのに対して，他のものは80%台に落ちる．

この傾向は単語認識率についても同様であり，その度合いは大きい．例えば，ExperVisionの単語認識率がGroup 1に対して99.8%，Group 5に対して87.7%であるのに対して，Caereは99.5%から77.5%に落ちる．60%台に落ちるものもある．

単語認識率と単語長の関係を見ると，上位のOCRは長い単語に対しても安定して高い精度を出すのに対して，下位のものは長い単語の精度は落ちる傾向にある．

CaleraとExperVisionは，ストップワード(2-5文字)に対して98%前後，非ストップワード(4-13文字)に対して96%前後を達成している．長い単語に対して精度が落ちないことから，単語照合を行っていると思われる．

興味深い事実は，文字認識率と単語認識率の関係が，ほぼ線形であることが分かったことである．通常，統計的独立を仮定して，単語誤り率は文字誤読率のベキ乗に比例すると仮定することが多いが，実際には異なる．これは，単語照合を行っていること，および誤読は低画質文書に集中するという性質に起因すると考えられる．

4．イメージ文書検索技術の動向

文書認識と情報検索の境界領域には3つの部分領域（研究アプローチ）がある．

1）情報検索に適した文書認識およびその後処理
2）文書認識(OCR)の利用に適した情報検索
3）文字認識を行わないイメージ文書の検索

以下，それぞれについて，SDAIRにおける主な動向を紹介する．

4.1 情報検索に適した文書認識

文書認識の目的が情報検索であることから，基本的には単語（非ストップワード）の認識精度が問題になるので，その精度を向上させることを主眼にする方法が多い．誤読や不読（読取不能／拒絶）をもたらす原因の多くは，文字の接触による切出し誤りにもあるという理由からも，単語単位に認識する方法が取られる．マルコフモデルやHMMを用いる単語認識はその例である [4, pp. 174-185] [7, pp. 203-216]．単語切出しを行わない単語認識の試みもある [6, pp. 177-188]．

特筆すべき研究として，文書中の単語イメージのクラスタリングを行って，同一単語のイメージと見做せる単語イメージを事前に抽出して，情報の冗長性を高めることによって，低画質文書の単語認識の精度を高めようとする新しいアイディアがある [4, pp. 26-39] [5, pp. 217-232] [7, pp. 177-190]．

しかしながら，一般には言語情報，特に語彙情報（シソーラス）を利用することが必要であるため，その不完全性が問題になる．この問題を回避する興味ある方法として，単語照合に，検索対象の文書データベース（コーパス）を用いる方法が提案されている [5, pp. 147-156]．完全に認識が終わっていない認識情報から，その文書に類似した文書をコーパスから検索して，それらに現れる単語の集合をシソーラスの代わりに単語照合に用いる方法である．

情報分類 (document clustering/document categorization) を目的とする場合は，検索条件が事前には未知な情報検索とは異なり，言語情報を利用する単語認識は有効である．

情報検索とは直接関係しないが，本会議で活発に議論されているテーマとして，人工的な低品質文書イメージの生成モデル (document degradation model/document image defect model) がある [6, pp. 127-136] [6, pp. 137-150] [7, pp. 217-228] [7, pp. 413-422]．高精度な文字認識を実現するためには膨大な量の変形文字イメージが必要であり，それらを自動的に生成する

のが第一の目的であった．しかし，OCRのエラー（誤読／不読）に影響されない情報検索手法の実験的研究のために，人工的な低画質文書データを供給する必要性が出てきた．

4.2 文書認識に適した情報検索

OCRエラーが情報検索に与える影響の程度に関する研究がある [6, pp. 115-126] [8, pp. 179-190]．OCRエラーを模擬的に生成して，それに対応するインデクスを作成し，再現率／適合率の評価をINQUERYシステムを用いて行っている．その結果，認識精度が高いときは，ほとんど影響がないが，精度が低くかつ短い文書を検索する場合は，かなり影響があることが分かった．ちなみに，情報検索における経験則 (rule of thumb) では，平均適合率の変化が5%以下の場合は「ほとんど影響がない」と見做される．一方，その変化が10%以上の時は「相当影響あり」と見做される．

別な研究では，ランキングに対する影響を特に調べている．この論文では，各種の検索手法を統一的な記法で比較評価している [8, pp. 255-270].

文書登録時のキーワード選択を単語イメージから行う研究がある [6, pp. 151-160]．単語イメージのクラスタリングを行って，同一単語と見做せるものの単語認識から，多くの単語仮説を生成し，simulated annealingの手法を用いてキーワードの選択を行う．この方法は，単語認識の高精度化と考えることも出来る．

OCRエラーが存在することを認めた上で，ベクトル空間モデルを用いて文書分類の精度を確保する方法の提案がある[7, pp. 301-316]．基本アイディアは単純で，誤読した単語文字列も含めてサンプル文書の学習を行う方法である．

OCRエラーの影響を軽減する方法として，ベクトル空間モデルを用いる場合の文書長の正規化方法の改善が提案されている [8, pp. 149-162]．従来，文書長の正規化のために，ベクトル間のcosineを検索条件と被検索文書との類似性の評価に用いる方法があるが，誤読の結果，出現頻度の少ないタームが出現して，この正規化法は悪影響を与える．その代わりとして，単純に文書長をバイト数によって正規化する方法を実験して，有効性を確かめている．この研究では，idf (inverse document frequency) を類似尺度に用いる場合は，元々cosineは用いるべきでないことも，大量の実験サンプルで示している．文書長をバイト数で正規化する場合は，idfを用いても良いとしている．

OCRエラーの影響を軽減する別の方法として，「ファジー検索」が提案されている [8, pp. 255-270]．従来の検索では文字列（ターム）間の完全一致が類似性の評価の基本になっていたが，（部分）文字列間の距離を「編集距離 (edit distance)」で計測し，0～1の値を取る類似尺度に変換する方法を提案している．更に，ブール論理検索，近傍検索にもファジー論理を適用している．有効性の評価は，OCRエラーのないクリーンなテキストを検索対象にしたときのランキングと，OCRエラーに汚されたテキストを検索したときのランキングとの差異の大きさで評価している．その結果，提案方式は従来方式より優れているとしている．しかし，再現率／適合率がどうなるかは示されてい

ない．また，検索処理の高速化が今後の課題である．

報告者らの研究として，認識候補の利用と混同行列情報 (confusion matrix) の利用がある [7, pp. 55-80]．日本語印刷文書イメージの検索に文書認識を用いた場合の誤読の影響を軽減する方式を２つ提案し，実験的に有効性を示した．

4.3 イメージ文書の検索

単語イメージを抽出し，従来の意味での文字認識／単語認識を行わず，単語イメージを符号化して，それを用いて文書分類あるいは情報検索を実現するアプローチが研究されている [5, pp. 105-122] [7, pp. 367-384] [8, pp. 163-178]．

単語イメージを矩形領域で切り出して，そのまま特徴抽出を行ってベクトル化する方法と，矩形領域に切り出された単語イメージを更に文字レベルに切り出して，それぞれの文字イメージを形状特徴から符号化する文字形状符号化 (character-shape coding) がある [7, pp. 367-384]．後者の場合，単語イメージは単語イメージトークン (word-shape token) に変換されることになり，以降は従来の方式が適用できる．

この分類のアプローチの研究は，まだ十分な規模の実験的評価はまだ与えられていない．

5．その他

その他の興味ある技術発表には，情報抽出のための自然言語処理を用いた属化抽出／標本法 (stratified sampling) [7, pp. 347-358]，コーパス依存の語幹処理法 [7, pp. 147-160]，あるいは階層ニューラルネットを用いたトピック抽出法 [8, pp. 317-332] が挙げられる．

応用システムとしては，必ずしも多くの発表はなかった．企業内メール配達の自動化を目的としたビジネスメール分類システム [5, pp. 443-456] [8, pp. 67-76]，米国特許庁文書データベースの紹介 [5, pp. 157-168] があった．

５回を通しての会議で，日本からの発表は，名大２件，日電１件，学術情報センタ１件，日立２件があった．

6．まとめ

今回のレビューでは研究動向が明らかになったのみならず，共通のデータベース／ツールなどの共用，問題提起を引き継いだ研究，関連研究の十分なレビュー，理解するに十分な長さの論文など，この会議 (SDAIR) の場が有効に生かされていることも実感した．メタなレベルでも学べる点があると思う．

参考文献

[1] UNLV: Information Science Research Institute 1993 Annual Reprot, April 1993.

[2] —: ISRI 1994 Annual Reprot, April 1994.

[3] —: ISRI 1995 Annual Reprot, April 1995.

[4] Proc. Symposium on Document Analysis and Information Retrieval, Las Vegas, March 16-18, 1992.

[5] Proc. Second SDAIR, Las Vegas, April 26-28, 1993.

[6] Proc. Third SDAIR, Las Vegas, April 11-13, 1994.

[7] Proc. Fourth SDAIR, Las Vegas, April 24-26, 1995.

[8] Proc. Fifth SDAIR, Las Vegas, April 15-17, 1996.