様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 ２６ 日現在

機関番号：１２６０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2021
課題番号：２０Ｋ２０６９２

研究課題名（和文）デーヴァナーガリー文字OCRの開発とサンスクリット文献データベースの構築

研究課題名（英文）Devanagari OCR and Sanskrit E-Text Archive

研究代表者
加藤　隆宏（KATO, Takahiro）

東京大学・大学院人文社会系研究科（文学部）・准教授

研究者番号：８０６３７９３４

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：本研究プロジェクトでは、AIエンジンによるデータ分析の材料となるデーヴァナーガリー文字の「字形データセット（教師データ）」作成を中心に行った。2021年7月には一度目のAI-OCRを生成して認識精度を検証した。その後もデータの追加とチューニングを繰り返し、最終的には1604文字種、48770文字数からなる字形データセットを完成した。このデータセットをもとに二度目のAI-OCRを生成し、サンプル文書を読み取って認識精度を検証した。この検証では、総文字数2434文字のところ96.14%（認識結果が正解文字のみの場合）、98.48%（認識結果の候補に正解文字が含まれる場合）という結果が得られた。

研究成果の学術的意義や社会的意義
本研究によって開発されたデーヴァナーガリー文字ＯＣＲは、第一の目的としてサンスクリット語文献（版本）をテキストデータ化するためのものであるが、その延長線上に開けた可能性として、インド国内外に大量に保存されているサンスクリット語写本資料をテキストデータ化への応用も視野に入れている。かつてマイクロフィルムに残されたものが、最近ではデジタル撮影・デジタルスキャンによって電子アーカイブ化が進められている。今後はこうした写本資料のテキストデータ化、さらには構造化が必要となってくるだろう。今回のＯＣＲ共同開発プロジェクトは、こうした研究の進展を見越したものである。

研究成果の概要（英文）：This project aims to develop a Devanagari OCR system. We introduced and evaluated preceding and on-going OCR software. We also reviewed the writing system of Devanagari and described how we correlate each combining letter with the Unicode encoding scheme. We took each letter as a composite of several elements. In this regard, we set a unit of letter called the “character shape.” We expounded the process of designing the “training data” through which an AI-OCR is generated. An AI-OCR was generated through machine learning using the prepared datasets. Following is a brief overview of the outcomes obtained from the generated AI-OCR model. Outcomes of Single Character Recognition (Out of the 2,434 sample letters): a. 2,340 letters exactly recognized (Accuracy rate 96.14 %) b. 2,397 letters correctly listed (Accuracy rate 98.48 %)

研究分野：インド哲学・サンスクリット文献学

キーワード：サンスクリット　OCR　デーヴァナーガリー

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

現代の文献学研究において、検索可能なテキストデータを用いた研究は欠かせない方法の一つとなっている。近年、サンスクリット文献学の分野で盛んに取り組まれている写本校訂研究においても、テクスト批評の方法論として、本文以外の関連文書の文体・文法・用例などを検討することの有用性は早くから認められ、検索可能なテキストデータベースの利用によって本文批評の手法は飛躍的に向上した。

しかしながら、既存のデータベースプロジェクトが提供するような、研究者それぞれの手入力によって作成されたデータベースには量的な限界があるのも事実であり、ある程度まとまった形のテキストデータベースを自動で構築するための文字認識技術の必要性がこれまで度々指摘されてきた。

２．研究の目的

このような状況をふまえ、本研究プロジェクトでは、サンスクリット語文献群を資料として用い、サンスクリット文献学を専門とする研究者とくずし字 AI-OCR 開発などを手がける凸版印刷株式会社の技術者との間で共同研究を行い、読み取り精度の高い OCR ソフトを開発することを第一の目的とした。以下、本報告では本プロジェクトで行った共同研究のうち、AI エンジンによるデータ分析の材料となる「字形データセット（教師データ）」を作成する過程で得られた諸課題とそれに対する考察と検討の結果、および生成された AI-OCR の読み取り結果についてまとめる。

３．研究の方法

３．１　ターゲット資料の選定

今回、OCR 生成のためのターゲット資料として、アーナンダ・アーシュラマ・サンスクリット・シリーズ（Ānandāśrama Sanskrit Series、現在 139 巻まで出ている）に収録された諸文献を用いた。同シリーズは、インド・プネー市にある出版元アーナンダ・アーシュラマが 1890 年代から 1930 年代頃までに出版した多くの重要サンスクリット文献を含んでいる。ここには神話・説話・文学・歴史・法典・科学・哲学等、あらゆるサンスクリット語著作がジャンルに偏らずに収録されており、多様な単語（＝多様な文字種）の採取が可能となると予想したからである。また、ジャンルの多様性は本研究プロジェクトのもう一つの柱である文献のデータベース化を行うのにも適していると言える。

３．２　転写方式

従来の研究ではデーヴァナーガリー文字を ISO 方式や KH（京都ハーヴァード）方式に基づきローマ字転写して記録したデータが広く用いられてきた。本研究ではデジタルデータ化の方法として、デーヴァナーガリー文字 Unicode に対応付ける方法を採った。デーヴァナーガリー文字をデータ化する際には比較的長く続く文節や単語をどこで区切るのかという困難があるが、デーヴァナーガリー文字 Unicode に一対一で対応づけることによって、この問題を解消することができると考えたからである。

また、ローマ字転写による表記はインドにおいて必ずしも一般的とは言えない。そのため、この方式は主にインド人研究者やインド諸語の使用者などの利便性にも考慮したものでもある。

３．３　結合文字

デーヴァナーガリー文字は子音文字 33 字（特殊な結合文字 kṣa, tra, jña の 3 字を加え 36 文字とすることもある）と、母音文字 14 字と、その他いくつかの記号により構成される、左書きの音素音節文字（アブギダ）である。これらの文字のうち、一つの文字の上下左右に別の文字やそれに由来する要素・記号が様々に結合することで多様な結合文字が形成される。

OCR を行うにあたり問題となるのは、(1) 数百種類以上存在する結合文字をどのように取り扱うか（ある程度の構成要素に分解するか、音節をまとめて 1 字とみなすか）、(2) 特殊な形で結合する短母音 i の記号をどのように処理するか、という二点である。

デーヴァナーガリーの結合文字は理論上は無限に作りうるが、現実には最多で 6 つの要素の結合に限定され、活版印刷用の活字ともなればパターンがある程度限定できる。そのため、完成した一音節を一矩形として音節単位の翻刻作業を行うのが妥当であると判断した。また、この矩形の設定方法は母音記号や子音 r の記号の正確な機械認識のためにも有効であると予測した。

３．４　字形データベースの開発

字形データベース（教師データ）の作成については、既存の文字認識ツールを複数テストしてこれらの弱点を分析することから始めた。

Google のデーヴァナーガリー文字 OCR の精度は数年前と比べて大きく上昇しサンスクリットも対象言語に挙げている。しかし、サンスクリットのテキストを読み取らせてみたところ、以下の 4 つのケースにおいて文字が正しく認識されないことがわかった。

(1) 結合文字が長大である場合、(2) 頭線（シローレーカー）上に付された鼻音記号（アヌスヴァーラ）の位置にヴァリアントがある場合、(3) 頭線の上下に分かれる母音記号で、頭線の上

下間に一定のずれがある場合、(4) 結合文字が改行により次行冒頭に飛ばされた場合、(5) 子音 r が頭線上に鉤状の子音記号によって示される場合、である。
(2)、(3) は近年の印刷物では稀有ではあるが、活版印刷による版本では頻繁に起こるものである。サンスクリットで書かれた重要著作の版本で質の良いものの多くがこのような活版印刷によるものであることを考慮すると、Google によるデーヴァナーガリー文字 OCR への改善点として、これら 2 点が特に重要な位置を占めることになるだろう。

次に、ind.senz による OCR ソフトウェアを検証する。ind.senz はデーヴァナーガリー文字用 OCR を、ヒンディー語、マラーティー語、サンスクリット語の言語別に提供している。今回のターゲット資料をサンスクリット語文献としたことに鑑みて、このうちのサンスクリット語用 OCR（SanskritOCR; http://www.indsenz.com/int/index.php?content=about）を検討対象とする。

テストページで発見された誤認識全 43 件のうち、文字の掠れに起因すると思われるものを除外すると、大きく次の 3 つの場合に誤認識が生じやすいことが判明した。

(1) 頭線（シローレーカー）上に付された鼻音記号（アヌスヴァーラ）の位置にヴァリアントがある場合、(2) 頭線の上下に分かれる母音記号で、頭線の上下間に一定のずれがある場合、(3) 特定の子音連続がある場合、である。

これらは上記で検討した Google による OCR にも共通するものであるが、SanskritOCR においては次のような特徴が見られた。まず (1) の場合には、存在する鼻音記号が認識されない場合のみならず、刊本には存在しない鼻音記号が認識される例も見られた。(2) については、母音 o、au、ɪ が ə に誤認識される（つまり、母音記号の頭線上部分が無視される）場合が殆どであった。最後に (3) については、子音連続の 2 文字目以降（例えば、sru という音節の子音 r）が認識されない場合や、dv などの特定の子音連続が他の単子音として誤認識される場合が多く見られた。

これら既存のツールの問題点をふまえ、字形データベースの作成時に、矩形（データ採取の際に四角形で囲む文字の最小単位）の範囲設定、翻刻・データ化の方法等について、文字システムや文法構造についての専門知識を活かしつつ検討した。また、採取した文字をもとに作成した出現頻度表を分析し、出現頻度の低い文字を選んで採取すると同時に、出現頻度が極端に低い文字種については対応する文字を完本の画像データをもとに合成してデータに追加することも試みた。これらの方法によって字形データベースの文字種・文字数がさらに充実した。

４．研究成果

４．１　1 文字 OCR の検証結果

以下、前節で述べた字形データベースをもとに生成した AI-OCR による 1 文字 OCR の検証結果を報告する。

・検証結果
総文字数：2,434 文字
(1) 認識結果の候補最上位が正解文字の場合のみを正解とする
（最上位に同数候補があった場合は誤認識とする）
正解文字数：2,340 文字
認識精度：96.14%
(2) 認識結果の候補に正解文字が含まれている場合に正解とする
正解文字数：2,397 文字
認識精度：98.48 %

４．２　評価

上記で得られた AI-OCR の検証結果を精査したところ誤認識が起きているケースとして次のような事例が見られた。

(1) 活字のサイズ・潰れ・印刷の不鮮明に由来するもの
(2) 鮮明な活字であっても紛らわしい字形をしているもの
(3)（ヴァリアントの）字形を採集できていなかったもの
(4) 同じ字形でも前後の文字の影響により異なる字形と認識されていると思しきもの

この中で、(1) については常に一定程度の混入を避けえないものであるが、このようなケースでも必ず認識失敗しているというわけでもなく、一部が欠けていても正しく認識するケースは散見された。他方、(2) と (3) については、さらなる字形の採取により教師データを増やしていく必要があり、今後の増補・調整が課題である。

また、総文字数 2,433 字中、656 文字含まれていた頭線上に記号が存在する文字については、(1) のケースが少なからず混入している中で 90.40%（593 文字）の認識成功率であった。また、事例数は限定されるが，上付きの母音記号と子音 r の鉤状の上付き記号が同時に付されるケース（2.4 に示した既存 OCR ツールでも読み取りに失敗することの多い字形）である 11 例では，81.82%（9 例）の認識成功率であった。また、母音記号 i を含む音節については 283 例中の 93.29%（264 例）の認識成功率であった。4 文字以上の長大な子音連結については今回の評価用データには含まれなかったが、採集済の字形に合致するものであれば、3 文字程度の結合子音も適切に認識していた。

(2)、(4) のケースについては、n-gram を用いた調整なども視野に入れる必要があり、これらは今後の課題となるだろう。

４．３　成果発表

2020 年度の研究成果は、第 127 回人文科学とコンピュータ研究会発表会において「デーヴァナーガリー文字 OCR の開発」と題して口頭発表を行い、その内容を共著論文（「デーヴァナーガリー文字 OCR の開発」加藤隆宏・友成有紀・谷口力光・大澤留次郎・藤巻聡・岡田崇・橋本江美, 『研究報告人文科学とコンピュータ』127/1, pp.1-4, 2021.）として公表した。

2021 年度に得られた研究成果については、2022 年 7 月に予定されている国際学会 Digital Humanities 2022 において発表予定である。

5．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

1．著者名 加藤隆宏・友成有紀・谷口力光・大澤留次郎・藤巻聡・岡田崇・橋本江美	4．巻 127/1
2．論文標題 デーヴァナーガリー文字OCRの開発	5．発行年 2021年
3．雑誌名 研究報告人文科学とコンピュータ	6．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 加藤隆宏
2．発表標題 デーヴァナーガリー文字OCRの開発
3．学会等名 第127回人文科学とコンピュータ研究会発表会
4．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

－

6．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

7．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

8．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関