日本がん登録協議会
第31回学術集会
ポスター演題 R-1-7

画像認識技術を用いた『外字』のユニコード文字探索支援

田渕　健
東京都立駒込病院／東京都がん登録室

【利益相反】

【学術集会名】
日本がん登録協議会第31回学術集会

【筆頭演者氏名】
田渕　健(東京都立駒込病院)

【COI開示】
当演題に関し，開示すべきCOIはありません

【前提】電子情報における文字情報

- 電子情報は言うまでもなくデジタル化されている
- 電子情報における文字情報はコンピュータ内部では文字コード(数値)として扱う
  - 英数文字は1バイト(256通り)で表現出来る
  - 漢字や仮名は少なくとも2バイト(65536=256²通り)で表現
- デスクワークの大半は既存の文字情報の操作
  - 新規創作することは少なく既存の雛形やデータの処理
  - 作業方法はコピーアンドペースト，検索・置換等
  - しかも処理内容の大半はルールベースド
  - コンピュータが処理すれば100万行程度であればほぼ一瞬

【背景】文字コードについて

- 1978年JIS第1･2水準(JISX0208の前身)制定
- 1980年住民基本台帳に漢字オンラインシステム導入
- 2000年JISX0213(JIS第3･4水準を含む)制定
  - 2000年e-Japan重点計画⇒汎用電子情報交換環境整備
  - 2003年JISX0213がunicodeに正式対応
- 2002年住民基本台帳ネットワーク稼働
  - JISX0221(国際符号化文字集合(UCS)対応)を独自拡張
- 2010-年文字情報基盤整備事業:文字情報基盤文字
  - 文字情報基盤漢字58712種(2019年5月版58843種)
  - 行政で用いられている文字を整理統合⇒相互運用性向上
  - 2013年Windows 8+Office 13でunicode完全対応

【背景】市区町村が使用する外字

- 「市区町村が使用する外字の実態調査」
  - 2011年度総務省(富士ゼロックス)
- 1,166,536外字(1,386市区町村:任意提出)
  - 自治体毎に独自外字を作成しているため膨大な数にのぼる
    - 電算化時期が異なる，当時の外字マッピング領域は全ての外字を網羅できず，地域の事情で作成する文字を選択せざるを得なかった
  - 2011年4月現在市区町村数合計は1,727⇒協力は約8割
- 文字情報基盤文字との対比
  - 2011年時点では変体仮名非実装

【背景】外部照合における外字の実態

- 当登録室で受けた都内2区市の外部照合用リスト(がん検診精度管理目的)はいずれも独自外字が用いられていた
- 他の区市町村も同様な実情とされ，外字使用実態報告書10年経過後でも外字使用の実態は変わっていないと言える
- 同報告書の指摘「システム間連携においても,外字データの連携のための同定作業や文字コードの変換テーブル作成の作業が発生し,作業負担や移行コストが増加」
- 外部照合を用いた情報の利活用の足枷になる可能性

【目的】

- 外部照合業務において，人手による判断が必要で最大の律速段階である外字同定を人手やコストをかけずに実現する

【方法】がん登録基盤業務としての位置づけ

- 本手法の開発は，がん登録個人情報は用いていないため，安全管理措置の適応外である
- 本手法で用いている外字フォントは提供元の基幹業務で用いられているが，外字そのものは公開されていないので行政機密に該当する可能性がある

⇒個々の外字について具体的に言及することは出来ない

- 本演題で述べる様な事象は一般的には体系的に扱われていないようであるが当登録室ではがん登録基盤業務として体系化している
- 本開発に要した余剰の人員及び資金はない

【方法】漢字正規化(Shift-JIS化)

- 全国がん登録で用いる漢字はShift-JIS漢字で表記
  - 実際にはCP932(拡張Shift-JIS)も使えるものが多い
  - 標記困難漢字は備考欄に言葉で説明との指示
    - ⇒鍼:JIS第3水準漢字⇒●:ゆみへんに前に刀(実はCP932漢字)
  - 実務者間の表現・解釈のゆれと自動処理の高い障壁
- 当登録室における漢字正規化手法
  - 漢字は「文字情報基盤MJ文字」(6万字弱)の範囲で表記
  - IPAによるMJ縮退マップ(JIS1-4水準漢字への対応表)
  - 異体字データベースを用いたShift-JIS漢字への置き換え
- 照合には，等価漢字(同一視文字)を定義
  - MJ縮退マップ，異体字と類似文字によって定義

【方法】文字情報基盤(MJ)文字

- 独立行政法人情報処理推進機構(IPA)文字情報基盤整備事業で整備された文字(漢字)
  - MJ文字情報一覧表Ver.006.01:58862文字
  - MJ文字情報一覧表変体仮名編 Ver.002.01:299文字
- MJ縮退マップ
  - MJ文字とJIS X 0213(JIS第1-4水準)文字との対応関係
- IPAmj明朝フォント
  - MJ文字情報一覧表掲載文字を表記(現行MJ文字は変体仮名を含む; 但，一部MJ文字は実装されていない)
- IPA(経産省政策実施機関)から一般社団法人文字情報技術促進協議会(CITPC) に民間移管(2020年)

【方法】外字フォント

- Unicode私用領域(外字領域)
  - E000～F8FF(6400文字)にマッピングされる
    - 現行の一般的なPC(Windows 8以降)で利用可能な領域
    - Shift-JIS以外の住基統一文字を全て実装するには不足するので，地域の実情に合わせて，文字を選定している模様
    - 補助私用領域A(FFF80～FFFFF),補助私用領域B(10FF80～10FFFF)は通常の環境では使われない
- EUDCフォント
  - Unicode私用領域に定義された外字をWindows-PCで外字を表記するためのフォント(環境依存:市区町村毎異なる)
- 登録室で受領するデータに外字が含まれている場合，EUDCフォントを併せて受領する

【方法】文字の類似度を判定する方法

- 現行用いられている外字のUnicode (或いは文字情報基盤(MJ)文字)へのマッピングルールは分かってない
  - ルールが分かっていれば，外字を利用している環境(区市町村や病院，企業等)では既にUnicodeに置き換えているはず
- 外字が用いられたデータを受領したら，登録室で外字とMJ文字との対応関係を見いだす必要がある
  - データが少なければ人手作業で行うことも可能
  - 資金と体制が整えばベンダーに委託することも可能
- 人が文字を判別するのは概ね文字の形で判定
  - 文字フォントを画像化して画像の類似度判定を行うアイデア
  - 人が文字を判定する場合は意味や部首の理解も伴う

【方法】文字フォントの画像化

- 対象の文字フォント及び文字コード
  - EUDCフォント:私用領域でフォントが実装された文字コード
  - IPAmj明朝フォント: MJ文字リストで定義された文字コード
- 各フォントを1文字ずつHTML表示し画像ファイルとして出力
  - 類似度判定効率向上のため，画像のサイズや解像度を調整
  - 手順はPythonのプログラミングにて作成
  - PythonのPILモジュールにおけるImageFont関数を用いるとフォントの画像化は容易に可能であるが，Unicodeの枝番を用いる形式のIVS/IVD文字は字形を忠実に出力できないという致命的な欠陥があるので使えない

【方法】画像化されたフォントの類似度

- 外字と文字情報基盤文字との対応づけの方法
- 特徴量マッピングによる類似度判定を行う
- フォントを画像化した画像データから『特徴量』を抽出
  - 特徴量マッチングは角や輝度の勾配などの特徴的な部分を抽出するアルゴリズムである
  - プログラミング言語PythonのライブラリOpenCV(cv2)にはAKAZEという特徴量検出器が実装されている
- 類似度判定アルゴリズム
  - 総当たりマッチング(Brute-Force matcher)
  - K-Nearest Neighbor algorithm(KNN)
  - いずれもcv2に実装

【結果】候補決定の可能性

- 類似度判定はマッチング精度Ratioやマッチングした特徴点の数に影響を受ける
  - 精度や特徴点の数を増加すると処理時間は大幅に延長する
- 完全字形一致文字＋デザイン差文字(併せて全体の約1/4)はマッチング精度を緩めに設定するRatio(>=0.99)ことで，ほぼ正解が得られた
- 類似文字(約半数)については，複数の候補を示し，目視選択が必要であった
- 同定不可能文字や記号(MJ文字に該当文字が存在しない)の類似度判定は意味をなさない
  - 外部照合リストには含まれないので実用的には問題がない

【結果】処理時間

- フォントの画像化に要する時間
  - EUDCフォントとIPAmj明朝フォント併せて6万字について30分程度で処理可能
- 類似度判定に要する時間
  - EUDCフォント1文字につきIPAmj明朝フォントを検索する時間は約2分半以内
  - EUDCフォント全ての検索を行うと，約4-5日要する
  - 実際のデータに出現するEUDCフォントは最大300文字程度であるため実際上は10時間程度で処理出来た

【考察】コンピュータリソースと漢字表記

- 現時点で，日本語は，中国語文化圏を除いて自国語表記に漢字を用いる唯一の言語である
- 日本では，中国よりもコンピュータ導入が早かったため，当時の貧弱なリソースで漢字を表記するため様々な工夫が必要であった(漢字ROM)
  - 第2次世界大戦後の漢字数制限の影響が残っており，コンピュータで用いる漢字は制限もやむを得ないという風潮
- コンピュータの急激な進歩で漢字表現力が高まり，多数の漢字が文字コードとして標準化されている
  - 中国の経済成長の急上昇とともに国際的な文字コード規格に中国の影響力が増してきたことも影響

【考察】漢字知識の共通認識

- 漢字知識の確固たる共通認識はない(個人差大きい)
  - 戦後漢字制限があったが最近は漢字表記も多様化
- 照合業務における目視同定は漢字知識について共通認識があることを前提を暗黙の了解としている
- 漢字知識は，院内がん登録実務者研修によるがん登録知識などとは異なり，無意識のうちに運用している
- 目視同定エラーの一部はこの共通認識のずれによるもので，一定比率は発生しうる
- 漢字表記ルールを全て書き出すことは一般には行わないので職務指導になじまない

【考察】人手作業による漢字置き換え

- 地域がん登録の手引き改訂第5版詳細版では第3水準漢字を第1-2水準に対応させる作業を例示で示している
- この対応に一般法則はないので，各登録室で人手作業による経験的作業を行う必要があるが，作業者間のゆれが大きくなる

⇒当登録室では文字情報基盤事業の知見と異体字データベースを組合せて網羅的リストを作成

【考察】人間とコンピュータの判断の違い

- コンピュータは自ら学習する(機械学習など)か否かは別としてルールと知識に基づいて動く
- 伝統的なコンピュータ運用方法は，コンピュータにルールと知識を事細かに指示する必要がある
- 我々が標準的だと思っている作業手順は，明文化されていても，コンピュータにとっては指示が不十分な情報
  - 職人の作業工程や料理のレシピ，がん登録標準業務手順
- われわれには，(時には無意識的に)前提知識を用いて直観的に判断している

【考察】JIS規格の字形・字体の変遷

- JIS規格のバージョンによって字形・字体が変遷
- JISX0208(概ね第1-2水準漢字)
  - 1978年制定，1983年改定，1990年改定
- JISX0213(第1-4水準漢字を含む)
  - 2000年制定，2004年改定
- 地名に関係が深い文字の変遷例
- 自治体の電算化時期によって外字対応が異なる
  - :JISX0208(1978)⇒:JISX0208(1983)⇒:JISX0213(2004)
  - :JISX0208(1978) ⇒:JISX0213(2004)
- 現在は，IVS/IVDで表記可能

【考察】提案手法の有用性と今後の課題

- 本手法では多くの文字で高精度で外字のUnicode文字を同定することが可能であった
- しかし，現時点では，完全自動で決定することは出来ず，人手による最終判断は必要である
- 人の判断のための支援技術としては有用であり，労力軽減と作業時間の短縮につながる
- 人間の作業は一定の比率でエラーが発生するので，エラー率を下げる効果もあると考える
- 今後の課題としては，深層学習による画像認識技術を用い，人間以上の判定精度を実現したい

【結語】

- 外字文字は馴染みが薄いため人手作業による類似判定は非常に骨の折れる仕事である
- 本提案手法は，人手作業を大幅に軽減する効果的な支援技術と言える