7ZC-04

漢字構造の自動判別とそれを用いた漢字入力方式について

金子　隆佐†　　相田　仁‡

東京大学†‡

1　はじめに

今日、日本語において漢字を入力する際に使用される入力方法の圧倒多数を占めるのは漢字の読みを利用したものである。しかしコンピュータで利用できる漢字はますます増えつつあり、2017 年 6 月に発行された Unicode 10.0 に収録された 9 万字弱に上る漢字（CJK 統合漢字）の大部分は日本語の読みが存在しない。そのためこれらの漢字は読みに頼って入力することが不可能である。また特に日本語においては同音異字が多いため、読みから単漢字を入力するのが困難な場合がある。

そこで本研究においては、「形・音・義」の漢字の 3 要素のうち形に着目し、漢字の構造を利用して漢字を入力可能なシステムの開発を目指し、その漢字構造を自動判別できるシステムの開発を行い、その評価を行った。

2　関連研究

かつては、読みではなく直接漢字等にコードを割り当て、それを入力する T コードなどの開発、研究が行われていた [1]。また中華圏においては倉頡輸入法や五筆字型といった入力法が現在も使用されている。しかしこれらは基本的に手動でコードを割り当てるために漢字の増加に迅速に対応できるものではない。

CHISE プロジェクト[2]は Unicode の IDC を用いた表現（例：砕→⿰石卒）で漢字構造の DB を作成しているが、編集は手動であり、また Unicode 以外の文字コードも参照しているため、漢字の増加に自動対応はせず、簡便に利用可能とはいえない。

A Study on Automatic Detection of Kanji Structure and Its Application to Kanji Input Methods
†Ryusuke KANEKO, The University of Tokyo
‡Hitoshi AIDA, The University of Tokyo

3　漢字構造判別システム

本研究では、後述する漢字入力システムで利用可能な漢字構造をコンピュータにおいて自動で判別するために以下のようなプログラムを作成した。

1. IDC（⿰⿱⿲⿳⿴⿵⿶⿷⿸⿹⿺⿻）を基にした漢字のタイプ分類
2. 1 の情報に基づいた漢字の分割
3. 分割した部分のマッチング（文字判定）

漢字のタイプ分類

漢字のフォント画像を 10 の漢字タイプ（上記の IDC から合成⿻を除く）に分類する畳み込みニューラルネットワークを作成した。CJK 統合漢字を全て網羅した花園明朝と、CHISE のデータをもととした訓練データを使用した。ただし、タイプ別の漢字分布に大きな偏りがあった（⿰:52,817 に対して⿶:50）ため、漢字を創作し不均衡を是正した。

テストデータで正解率を検証したところ、⿱と⿳というような IDS 表現の任意性によるミスが目立ったため、⿱と⿳、⿰と⿲を同一のものと縮退したところ、タイプ別では図 1 に示したように平均正解率は 98.2%となった。

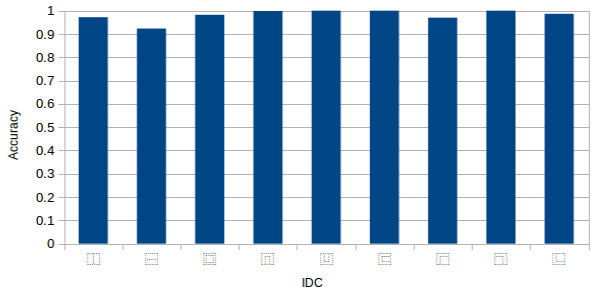

図 1　漢字タイプ別の分類正解率

漢字分割

前小節で得た情報（すなわちどの方向に漢字を切断できるか）をもとに漢字を自動分割する。[3]などが示唆するように最も単純な方法はある方向を走査してそのうち黒画素が最も少ない部分で分割する

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.

というものである。しかし、漢字のフォント画像分割においては、必ずしも黒画素の少ない部分が切断に適切ではないという問題等がある。本研究においては、走査線が通過する黒画素からさらにその垂直方向にどれだけ黒画素が続くかを計量することで、画素は少ないものの切断に適さない横画の部分などでの分割を抑制することを試みた。図 2 にその例を示す。

図 2　黒画素の最小な線で分割した例（左）と提案手法により分割した例（右）(U+22A9A)

これにより分割の正解率は 59.2%から 74.1%へ向上した。

分割部分のマッチング（文字判定）

分割した漢字の構成部品をテンプレート群とマッチングすることで、その文字を特定する。[4] では白黒 2 値画像のための様々な類似度関数の評価が行われているが、本研究ではこの中から全般的に好成績を収めている YULE の手法を利用することとした。

ただし漢字タイプによってはその判定精度が大きく劣る場合があった。例えば上述のタイプ⿴にあたる「⿴尺」(U+2C936)の尺部分だけを切り取った歪んだ画像と正方形に近い「尺」とを比較した場合などである。このために他の類似度計算手法として、近傍点距離（Nearest Neighbor Distance）を定義しこれによって類似度の計算を行った。すなわちテンプレート画像の黒点とそれに対応するターゲット画像の最も近い黒点との距離d_{NN}を定義し、すべての黒画素について、次のようにスコアを算出する。

$$\mathrm{Score}_{\mathrm{NND}} = \sum_{\mathrm{black}} \exp[\alpha d_{\mathrm{NN}}]$$

係数αは黒点が離れていることへのペナルティ係数である。これを用いて類似度判定を行い、YULE 法と比較した結果、一般的漢字については 0.2 だけ正答の平均順位が悪化したが、タイプ⿴では 2 だけ平均順位が改善した。

作成したシステム

以上の手法を総合して作成したプログラムは Computer-Aided なものにとどまったものの、任意のテンプレート群の文字に漢字を分解できるシステムを開発した。

4　漢字入力システム

提案する漢字入力システムは、例えば「勛」(U+52DB)という漢字を「員」と「力」からなっていると判断し、「員、力」というふうに入力するとこの 2 つの漢字を構成要素に含む漢字が得られるというものである。

有効性を示すためにここでは、中華、朝鮮圏の人名に使用される漢字であり本邦の常用漢字や人名用漢字に含まれない漢字群を用意し、特定の漢字を入力するのに必要な最低打鍵数（ローマ字と候補選択に必要なスペース等の打鍵回数）を比較した。結果を表 1 と図 3 に示す。なおこの 200 字中 2 字については日本語読みが存在しない漢字だったが、提案手法により間接的に読みを使用して入力可能となった。

表 1　漢字入力システムの打鍵数による評価

テスト群の漢字数	200
うち打鍵が減少した漢字数	92
打鍵減少度の平均	14.7

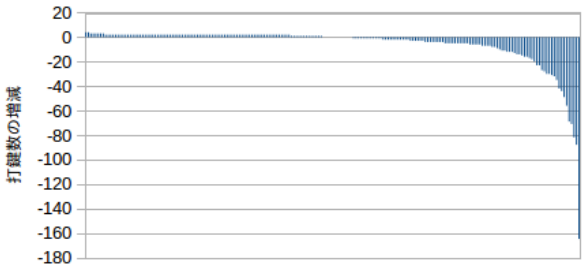

図 3　漢字ごとの打鍵数増減幅の分布

5　まとめ

ますます増えつつある漢字の入力に資する一連のシステム開発を行った。今後は自動判定の精度向上を図りたい。

参考文献

[1] 山田尚勇「コンピュータ科学者がみた日本語入力と入力 2：文字入力とテクノロジー」，くろしお出版，2014

[2] Morioka, T. “Multiple-policy Character Annotation based on CHISE”, Journal of the Japanese Association for Digital Humanities, 2015, 1, 86–106

[3] Casey, R. G. and Lecolinet, E. “A survey of methods and strategies in character segmentation”, IEEE Transaction on Pattern Analysis and Machine Intelligence, 1996, 18, 690–706

[4] Tubbs, J. “A note on binary template matching”, Pattern Recognition, 1989, 22, 359–365

Copyright ©2018 Information Processing Society of Japan.
All Rights Reserved.