国立大学法人 東京農工大学

NEWS RELEASE

報道関係者　各位

2017年12月11日
国立大学法人　東京農工大学

古典籍のくずし字をAIが認識

国立大学法人東京農工大学大学院工学研究院の中川研究室の修士課程2年リー・トゥアン・ナムと博士課程1年グエン・コング・カーは、電子情報通信学会　パターン認識・メディア理解研究会（PRMU）が主催する「第21回アルゴリズムコンテスト（注１）」において、難易度の高いレベル2とレベル3（注２）で優秀な認識性能を示し、コンテストに応募されたアルゴリズムのうち、最も優秀と判定されたアルゴリズムを考案した者1件に贈呈される「PRMUアルゴリズムコンテスト最優秀賞」を受賞しました。これに先立ち、11月に開催された4th International Workshop on Historical Document Imaging and Processing (HIP 2017)でも、最優秀論文賞を受賞しております。

電子情報通信学会　パターン認識・メディア理解研究会（PRMU）第21回アルゴリズムコンテスト表彰式
2017年12月16日(土)
※報道解禁：12月17日(日)0時

各レベルにおける本学チームの結果
・レベル2
- 認識率：87.6%（23チーム中1位）
- 認識時間：一文字あたり2.16秒（23チーム中10位）

・レベル3
- 認識率：39.1%（23チーム中1位）
- 認識時間：一文字あたり0.43秒（23チーム中5位）

講評
1点から5点の採点で、3人の評価者から平均して、新規性4点、信頼性5点、明瞭さ4.33を得ました。明確な文字の切出しを必要としないこと、種々の構成を検討していること、レベル 3 のための複数行の検出と統合、総合的な構成が評価されています。

方式
レベル 2 のためには、これまでに提案されているいくつかのニューラルネットワークを検討し、Convolutional Neural Network（CNN: 畳込みニューラルネットワーク）、Bidirectional Long Short-term Neural Network（BLSTM: 双方向長・短期記憶ニューラルネットワーク）、そして, Connectionist Temporal Classification（CTC: コネクショニスト時系列識別法）を 3 層に組み合わせ、Deep Convolutional Recurrent Network（DCRN: 深層畳込み再帰ネットワーク）を構成しています（図 1）。第一層では、事前に学習させたCNNによって縦書きのくずし字から特徴の列を抽出し、第二番目の再帰層ではBLSTMによって候補文字と確率の組の列に変換し、3番目の層のCTCで文字列に変換します。

レベル3では、X-Yカット法とボロノイダイアグラムを使って文字行を切り出し、行を一列につないでから、レベル2の方式を適用しています。X-Yカット法は、縦方向、あるいは、横方向への射影によって空白を見つけて切り、次に、他の方向への射影において空白を見つけて切るという作業を、切れなくなるまで繰り返す手法です。ボロノイダイアグラムは、複数の黒画領域に等距離の白画素内の点をつないでできる分割図です。斜めや複雑に空間がある場合でも切り出せるというメリットがあり、X-Yカット法で分離できないものに、ボロノイダイアグラムを利用します。

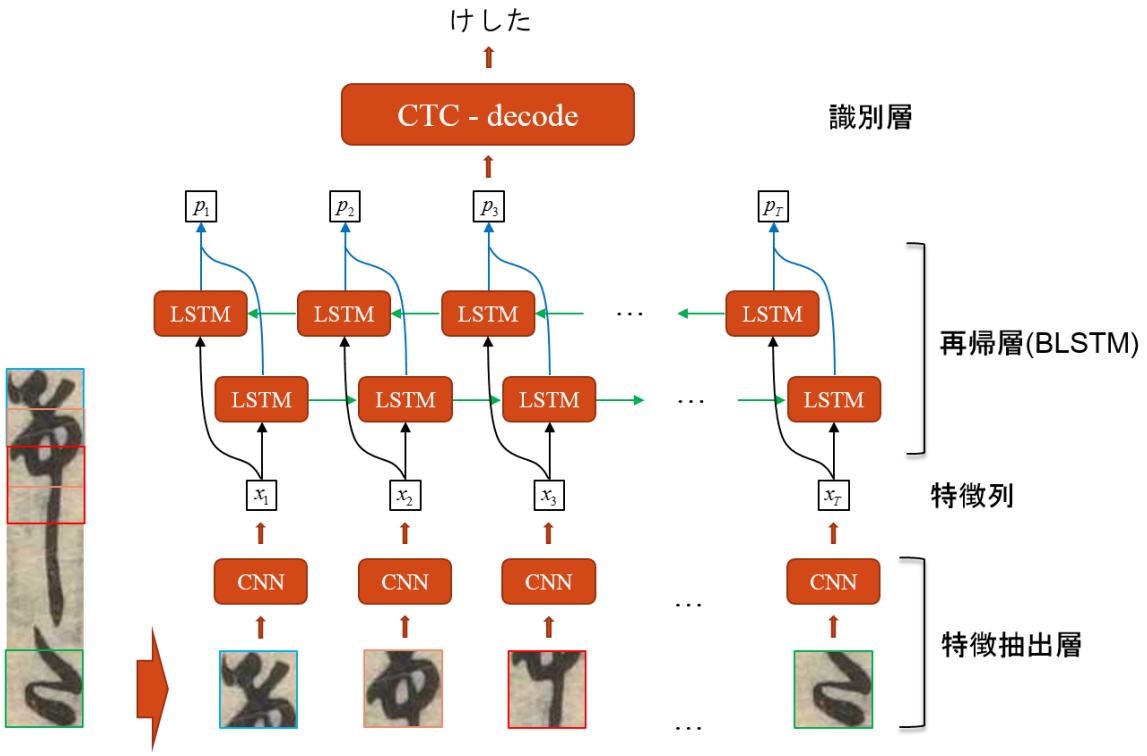

図 1：DCRN の構成

今後の展開

古典を電子的に保存・公開し、手軽に分析できるようにすることは、歴史研究には不可欠であり、古典籍で使用されるくずし字を正確に認識する技術開発を進めることで、歴史研究の発展に寄与できると考えられます。

今回、本学チームのレベル 3 の認識率が低かったのは、1 文字でも誤認識すると、全体として誤認識になるためです。文字列のなかに混同しやすい文字があると、こうした結果になります。これを改善するためには、当時の言語統計から文字と文字のつながりやすさ（文脈）を利用する方法が効果的であり、それによってレベル 3 の認識率はレベル 2 に近くなることが予想されます。データをさらに大量に蓄積することで、漢字を含めて、認識率はさらに高まることが期待されます。

（注 1）アルゴリズムコンテストの概要：パターン認識・メディア理解分野の若手研究者・学生の育成および研究会活動の活性化を目的として、毎年開催されています。提示される課題には、代表的・基礎的な研究課題が取り上げられ、応募されたアルゴリズムは、その性能・独創性・処理時間の観点で評価されます。（詳細：https://sites.google.com/view/alcon2017prmu）

（注 2）第 21 回アルゴリズムコンテストの課題とレベル：図 2～4 のように、古典籍画像の指定領域に含まれるくずし字を認識して、コードを出力します。

1. **課題の難易度**

   外接する長方形に含まれる文字数に応じて課題の難易度を設定しています。レベル 1 は 1 文字、レベル 2 は縦方向の 3 文字、レベル 3 は縦横方向の 3 文字以上の文字を含んでいます。

2. **認識対象の文字**

   認識対象の文字は変体かな 50 種程度です。漢字は含みません。

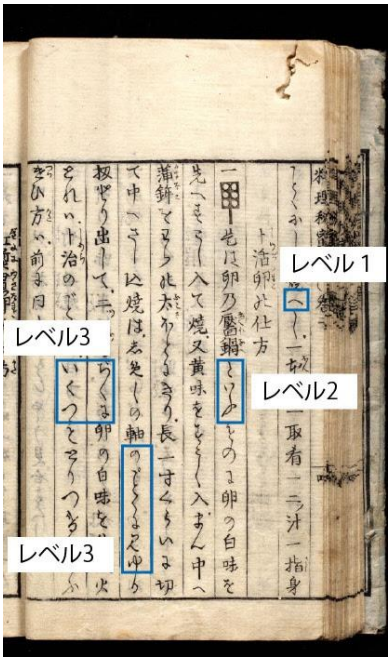

図 2：課題の 3 レベル。レベル 1 は 1 文字、レベル 2 は縦方向の 3 文字、レベル 3 は縦横方向の 3 文字以上。

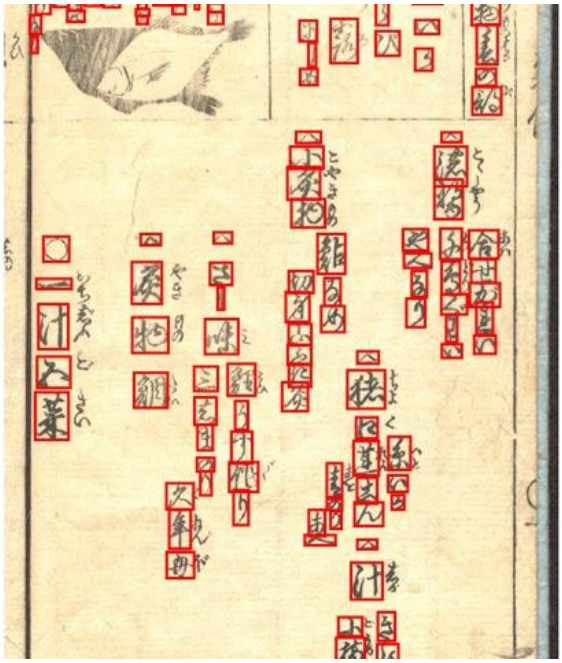

図 3：人文学オープンデータ共同利用センターが公開している日本古典籍字形データセット

異なる字種にも関わらず似た字形　　同一文字の異なる字形

図 4：人間でも読みにくい文字の組

◆ 研究に関する問い合わせ◆
東京農工大学大学院工学研究院
先端情報科学部門　教授
中川　正樹（なかがわ　まさき）
TEL/FAX：042−388−7144