令和元（２０１９）年度　卒業研究報告書概要

課程，学籍番号，氏名	課程：電気・電子情報工学課程，学籍番号：B173215，氏名：金井　達哉
工学分野名：情報通信システム	指導教員名：市川　周一

題　　目：和　近代及び近世の文献における２文字連接特徴とその利用に関する検討

（英　Research on the use of 2-gram in modern and early modern documents　）

Abstract

Though character inference is important to read Japanese old documents, it is difficult for ordinary people. A 2-gram is a contiguous sequence of two characters. The frequency characteristics of 2-gram is useful to infer the pervious or the next character (Yamada and Shibayama, 2003). This paper presents the results of cluster analysis by Euclidean distance using 2-gram in modern and early modern documents. The characteristics of 2-gram are influenced by writer and genre. Average of front and back character inference probability is 0.13±0.06. If multiple candidates were considered according to the frequency of front and back side, the characters could be proposed with an average accuracy of 40% to 60%.

概　　要

古文書を活字に直すことを翻刻という．古文書を理解したりその情報を活用するうえで重要な作業であるが，古文書は多くの場合くずし字で書かれている．くずし字では，複数の文字が同じあるいは極めて近い形になる場合がある．そのため一意に文字を判別できない場合がある．この場合，文脈や前後の文字から判断する必要がある．これには多くの経験や知識が必要となるため，一般利用者の古文書読解における一つの壁となっている．

ある文字の前後に来る文字の頻度や確率を連接特徴といい，英字の置換暗号の解読などで用いられている（松井（2001））．英語の場合は連接頻度や連接度数の表から，不明な文字や単語を推定することができることが知られている．本研究では，連接特徴の情報から不明文字の推測ができれば古文書読解の学習を支援できると考えた．

古典籍における文字予測としては，証文類を対象とした山田と柴山（2003）のシステムや，変体仮名を対象とした渡辺ら（2005）のシステムがある．本研究では公文書以外の文献における漢字と仮名の両方における連接特徴の調査を行った．

まず，近代と近世の文献のそれぞれにおいて 2 文字連接（2-gram）の出現率によるユークリッド距離でのクラスタリングを行った．その結果，2 文字連接特徴はジャンルや作者の影響を受けることを確認した．

次に，近世の文献 6 種類においての連接頻度を調査した．連接頻度は前連接頻度，後連接頻度ともに平均で 0.13±0.06 であった．また，連接頻度同士の積を文字予測正解率と仮定した時の正解率は，最大でもおよそ 10%程度であった．前後連接頻度のおおきいもの上位から順に複数候補提案した場合は，平均 40%～60%の精度で不明文字を提案できることが分かった．

また，山田と柴山（2003）の前連接頻度と後連接頻度を比較して大きい方を提案する手法についても上記と同様の実験を行った．その結果，一文字だけ提案する場合，平均連接確率は 0.33±0.1 だった．一方，複数候補提案した場合では，平均 30%～50%の精度で不明文字を提案できることが分かった．

提案精度の向上は今後の課題であり，頻度不明の文字連接を減らすためにコーパスを増やす事や，2 文字以上の連接特徴を組み合わせることが必要であると考えられる．

発表する際の課程を記入　　電気・電子情報工学　課程　　発表番号　14

（学籍が他課程所属の学生も発表する課程を記入すること）