二値分類と距離学習を組み合わせた Human-in-the-loop 書誌同定手法

大沢 直史

現在，日本では年間およそ 7 万冊の書籍が発行され，戦後から数えるとおよそ 300 万冊近くの書籍が発行されている．この書籍 1 冊 1 冊に与えられるタイトル，著者等のメタデータ(以下，書誌データ)は 1 点の書誌に対して大きく 4 種類，細かく分類すると数百種類に分類され，存在している．大きな 4 種類とは流通用のデータ，公立図書館のデータ，国立国会図書館のデータ，大学図書館のデータである．国立国会図書館と大学図書館のデータは 1 種類に定まるが，流通用データや公立図書館のデータはそれぞれ複数の書誌データが存在する場合がある．具体的にはタイトルの記述方法が違うことや著者の表記の揺れなどがある．そのため，図書館毎に作成される書誌データに差分が生じ，結果として 1 冊の書籍に対して複数の書誌データが存在することになる．この状況は，複数の図書館にまたがって書籍を検索することのできる横断検索システムにおいて，同一の書誌が違う書誌であると判断され，検索結果に書誌が重複して表示される「書誌割れ」や，異なる書誌が同一の書誌であると判断され検索結果に書誌が 1 冊に定まり表示される「書誌誤同定」を引き起こす．

書籍出版物の書誌を特定することができる ISBN は 1 冊の書籍に対して固有の番号が付与される．そのため，本来ならば ISBN が一致している書誌は同一の書誌であるといえるが，ISBN の取得には費用が掛かることなどから，市場に出ることのなくなった書籍の ISBN を別な書籍に流用する，いわゆる使い回しの事例も存在する．また，ISBN が国内で使用されるようになったのは 1981 年頃からであり，それ以前の書籍には付与されていない．これらより，ISBN を用いた書誌の同定は困難であるといえる．

これまで様々なアプローチでの書誌同定手法が提案されてきた．その中には機械学習等の自動化手法に関する研究も存在するが，必ずしもその精度に満足がいくものではなかった．一方，機械学習の発展は近年めざましく，より高性能な機械学習手法が現れている．また，人間の力を組み込んだ Human-in-the-loop アプローチも注目されている．

そこで本論文では，最新の機械学習の技術を適用して，Human-in-the-loop による書誌同定の自動化手法がどの程度の性能を発揮できるかを検証する．具体的には，深層学習を用いた二値分類と距離学習の利用について検証した．深層学習については書誌のペアを与えてそれが同一かどうかの判定にどれぐらいの性能がでるか，距離学習については判定が候補ペアの絞り込みに利用可能か，について検証を行った．

実験の結果，国立国会図書館の書誌データのうち ISBN を持つデータに関して，書誌が一致しているか否かを判定する二値分類モデルでは 87%の正解率を得られた．距離学習の利用に関しては，候補ペアの絞り込みや，人間が行うタスクの選択には有効ではあるものの，絞り込みに要する時間が距離学習によって得られる書誌データの次元数に大きく依存する．高速化のため，絞り込み精度を保った最低限の次元数を獲得する必要があるという課題が明らかになった．

（指導教員　森嶋厚行）