（別紙様式２）

公立小松大学重点研究「みらい」 研究実績報告書

氏名	所属・職名	助成金額
坂本一磨	生産システム科学部　助教	1,000,000 円

研究課題名	Web データと SNS の投稿を用いて自動生成した文章を活用したユーザの属性推定に関する研究
研究期間	令和3年　6月　1日～令和5年　3月31日
研究の概要	我が国では,「第 5 期科学技術基本計画」にてサイバー空間とフィジカル空間を融合させた新たな社会像である Society 5.0 が提唱され，ネットワーク技術や IoT（Internet of Things），そして AI（Artificial Intelligence）技術を用いて，経済成長に向けた社会課題を解決するための取り組みがなされつつある．その中でも，インターネットから社会の動向やニーズを把握することを目的としたソーシャルセンシング技術の研究が進められており，社会調査やマーケティング，データマイニング等の分野で活用されている．その中でも SNS（Social Networking Service）は，サイバー空間とフィジカル空間を融合したきめ細やかな情報交換手段として有効であることが確認されている．しかし，投稿記事の確からしさが保証されていないため，重要性が高くしかも信頼性も担保された情報を随時正確に抽出することは困難である．また，信頼性を担保する手段の一つとして，SNS などのサイバー空間の情報から，ユーザの性別，年代（年齢）や職業，そして地域などの属性情報を推定し，尚且つユーザの投稿履歴や URL などから個人の特徴を推測する取り組みが行われている．しかし，個人情報保護の観点から取得できる情報量が少なく，ユーザに適材適所のデータを提供することが難しい．そこで，Web 上から収集したデータを深層学習で解析し，各ユーザ属性と特徴となる文章を自動生成して，その内容からユーザ属性を推定する．また，情報は時々刻々と変化しているため，Web 上から情報を自動的に取得し，推定結果に応じて学習モデルを精査し，学習モデルが逐次拡張する仕組みを開発する．これにより，日々変化する情報に対応でき，データ量が少ない場合でも信頼性のあるユーザ属性を推定でき，対象にしたいユーザに合った情報を提供が可能となる． 本研究は，2 年間で，Web 上から収集した内容からユーザ属性ごとの文章を自動生成し，その文章からユーザの属性を推定する．1 年目は，Web 上から情報を自動的に取得するシステムと深層学習を用いて，文章を自動生成するシステムを開発する．2 年目は，生成した文章の特徴量から深層学習を用いて解析するシステムを開発し，ユーザの属性推定と高精度化を行う．具体的な方法を以下に詳述する． Web 上から取得するシステムでは，まず，モデルの基本となる SNS ユーザの属性を手動により付与して，ユーザの投稿時間と投稿内容を取得する．次に，各年度の出来事を Wikipedia から取得する．そして，ユーザの投稿内容と投稿時間から生活習慣を解析し，習慣的な行動を抽出する．さらに，習慣行動ごとに各年度の出来事に対する投稿内容をトピックに分類する．最後に，ユーザの投稿内容とトピックの内容から得られた単語，キーワードを用いて文章生成するためのデータとして抽出する．抽出したデータとユーザの投稿内容を深層学習の SeqGAN を用いて，ユーザの投稿内容を自動生成する．これによって，各年度で使用傾向が多い単語と各属性によって特徴となる単語を考慮した文章を生成することが可能となり，日々変化する情報や属性ごとに異なる文章にも対応できる． ユーザの属性推定では，自動生成した文章と SNS から取得した文章を深層学習の RNN（Recurrent Neural Network）の一つである Bi-LSTM（Bi-directional Long Short-Term Memory）を用いて，学習する．その際，自動生成した文章と SNS から取得した文章の使用割合による属性推定精度のパラメータ検証を行い，最適な学習モデルの生成に取り組む．本研究で対象とする属性は，職業（社会人，主婦，学生，フリータ），年代（10 代，20 代，30 代，40 代以上），地域（8 地方）とする．性別に関しては，既存技術において 9 割以上の精度で推定できていることから対象外とする．これらの内容から単語の使用方法や関係性を学習したモデルを構築し，各ユーザ属性を推定する．ユーザの属性推定精度の高度化に関しては，各ユーザ属性に暗黙的に存在する習慣行動の関係性を重みとして使用し，高精度化できるかを検証する．これにより，各属性の文章特徴を考慮してユーザ属性を高精度に推定でき，ユーザに合った情報を提供する技術に寄与できる．

<table>
<tr><td>研究の成果</td><td>当初の目的は，Web 上から収集した内容からユーザ属性ごとの文章を自動生成し，その文章からユーザの属性を推定することである.
1 年目は，自動文章を自動生成することで，ユーザの属性推定を行うシステムを構築した.
1 つ目の内容としては，Web 上から取集した文章から深層学習を用いて，男性と女性を対象に投稿文章の自動生成を行った．そして，自動生成した文章と実際に投稿した文章でのユーザ属性の推定を行い，評価した.
2 つ目の内容としては，ユーザ属性の中でも興味，関心に着目し，深層学習を用いて，ユーザ属性の推定を行い，評価した.
以上の内容を情報処理学会にて，学会発表（研究成果発表：（1）と（2））した．また，ユーザ属性の推定に関する内容の関連研究の成果として，学術誌掲載論文（研究成果発表：(1)）を成果として報告した.
2 年目の前半は，ユーザの属性推定の高精度化を目指し，研究を行っていたが，後半は，ユーザに合った情報を提供する技術の確立を主軸に研究を行った．具体的な成果としては，開発したシステムのさらなる汎用性と有用性の評価のため，リアルタイムに投稿されている文書を解析し，スポーツの野球プレーを推定するシステムを開発した．これらの研究成果を情報処理学会にて，学会発表（研究成果発表：（3）と（4）），学術誌掲載論文（研究成果発表：(2)）を成果として報告した.</td></tr>
<tr><td>研究成果発表状況</td><td>【学術誌掲載論文】
(1) 坂本一磨，中村健二，山本雄平，田中成典：マイクロブログユーザの類語に着目した地域属性の推定に関する研究，情報知識学会誌，情報知識学会，Vol.32，No.1，pp.53-72，2022.（査読付）
(2) Kazuma, S.，Riku，Ikeda.，Yoshihiro，Ueda.：Development of News Bulletin System from Real-time Tweets using BERT，Studies in Science and Technology，2023.（査読付）（投稿中）

【学会発表】
(1) 中川嵩将，上田芳弘，坂本一磨：BERT を用いたマイクロブログユーザの興味推定に関する研究，第 84 回全国大会講演論文集，情報処理学会，2022.
(2) 三村亮，上田芳弘，坂本一磨：SeqGAN を用いたマイクロブログの文章自動生成に関する研究，第 84 回全国大会講演論文集，情報処理学会，2022.
(3) 堂前拓生，上田芳弘，坂本一磨，池田理玖：SNS のテキストデータを用いた BERT による投稿者の属性推定，第 85 回全国大会講演論文集，情報処理学会，2023.
(4) 池田理玖，上田芳弘，坂本一磨：BERT を用いたリアルタイム投稿による速報システムの開発，第 85 回全国大会講演論文集，情報処理学会，2023.（学生奨励賞受賞）</td></tr>
<tr><td>経費の執行状況</td><td colspan="2"><table>
<tr><th>区　　分</th><th>執行額（円）</th><th>備　　考</th></tr>
<tr><td>カスタム PC</td><td>436,920</td><td>深層学習を用いた解析に使用</td></tr>
<tr><td>カスタム PC</td><td>498,300</td><td>深層学習を用いた解析に使用</td></tr>
<tr><td>論文印刷費</td><td>81,887</td><td>学会に投稿した論文の掲載費，印刷費に使用</td></tr>
<tr><td>BTO パソコン</td><td>60,880</td><td>学習データのクロール用に使用</td></tr>
<tr><td>ハブドッキングステーション</td><td>3,144</td><td>解析時に学生の PC も使用するため

※不足分は，別研究費から執行</td></tr>
</table></td></tr>
</table>