様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月　 ９ 日現在

機関番号：１６１０１
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１５４９

研究課題名（和文）言語と非言語の混在するWWW上の生活習慣・健康情報の統合的解析

研究課題名（英文）Analysis of lifestyle and health texts on the WWW consisting of text and numeric data.

研究代表者
吉田　稔（YOSHIDA, Minoru）

徳島大学・大学院社会産業理工学研究部（理工学域）・講師

研究者番号：４０３６１６８８

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：数値的情報と言語表現との関連を分析し、数値情報を言語表現に変換する手法に応用することで、血圧に関する文書等、Web上の健康に関する科学的情報へのアクセスを向上させるための研究を行った。テキスト中に記述された数値情報に関しては、適切な数値範囲の推定、および、その関連する言語表現の抽出技術を開発した。また、SNSへの投稿時刻をもとに推測された疑似的な睡眠時間をもとに、睡眠と言語的表現の関連についても分析を行った。
また、SNS上での食事画像と、それに付随するテキストを組み合わせたデータセットを、食事カテゴリ毎に作成し、画像的特徴量と言語的特徴量の関連を分析する基盤を構築した。

研究成果の学術的意義や社会的意義
数値表現と言語表現、あるいは、画像と言語表現の関連を分析し、相互変換を行うことで、従来実現できなかった、「数値表現をもとに、適切な文書を検索する」というシステムの実現が可能になった。また、SNS上でのユーザーの健康に関する文書を分析するための基盤を構築したことで、公的機関等による文書と、一般ユーザーの健康に対する報告を結びつける手法の実現に向けての環境が整った。

研究成果の概要（英文）：We investigated the relations between numerical information and linguistic expressions and proposed a method to convert numeric data into linguistic expressions that can contribute to accessing the scientific documents about healthcare. We considered two situations: one is the case where the numeric data were provided in text, and the other is the case where they were provided as metadata, e.g., the posting time of tweets.
We also developed the data sets that consist of (text, image) pairs from Twitter by searching food-related keywords for the purpose of analyzing the relations between food images and associated text. Our data sets can contribute to analyzing eating habits of people by complementarily using text and images about users' eating reports.

研究分野：テキストマイニング

キーワード：テキスト中の数値表現　健康情報検索

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

WWW の普及により、医療や健康に関する記事が、大量に流通しており、これら膨大な医療・健康情報への効率的アクセスの提供が、近年の情報爆発時代における喫緊の課題の一つとなっている。

体重や血圧といった体調に関する情報は、様々な病気と関連し、多様な医療関連文書に登場するため、「関連する病名で検索する」という通常の検索ではなく、「体調から関連する文書を検索する」という逆引きの手法が必要となる。

従来の研究は、キーワードの共起頻度を計測し、関連語を抽出する処理が一般的である。しかしながら、体重や血圧等、体調に関する客観的情報は、「~kg」といった具体的数値で報告されることが多く、また、その数値の表現も、「~kg 増えた」「~kg になった」「～kg を目指す」等多様であり、正確に意味を把握することは難しい。

一方、医療関連文書では、体調に関する記述は、「肥満」「睡眠不足」「高血圧」といった、言葉で表現されていることが一般的である。このため、「言語で記述されないユーザーの体調情報」を言語化し、言語で記述された医療・健康文書へ関連づける必要がある。本研究では、このような言語化を「タグ付け」と呼び、ユーザーが検索したい状態（体重や血圧等）を数値的に表現したものに自動的に「タグ付け」を行う手法について研究する。

２．研究の目的

本研究では、WWW 上に存在する「大量の医療関連文書」を、ユーザーの「体調情報」および「生活習慣」に関連づけるための手法について研究を行う。「病気」と「体調」の関連のみならず、「体調」と「生活習慣」の関連にも着目し、病気予防を、「生活習慣の改善」という、より具体的な方策に結びつけることを目指す。

生活習慣や健康情報は、SNS 上においては「体重 90kg」のような数値や、食事写真のような画像で表現されることが多く、WWW 文書においては、「肥満」等の言語で表現されることが多い。本研究提案では、これらの非言語情報と言語情報の混在した SNS の投稿や WWW 文書に関して、数値・画像と言語の間の変換手法を通して、統合的解析を提案する。また、WWW 文書から適切な部分文書を取り出すため、汎用性の高い高速なレイアウト解析手法を開発する。

３．研究の方法

本研究は、上記「体調」「生活習慣」の解析において、特に、数値や画像と言語の関係に着目している。数値と言語の関係について、分散表現やトピックモデル等複数の手法を拡張し、数値情報を扱えるようにすることを目指す。これにより、「体重」「睡眠」「血圧」等、数値的な値を持つ情報について、関連する言語的表現を獲得することを目指す。「体重」や「血圧」については、テキスト中に直接記述された数値を用いる。このさい、例えば「体重」に関して、「体重そのもの」か「減少幅」かを分類することも検討する。これに対し、「睡眠」については、値（睡眠時間）がテキスト中に明示されることは少ないため、テキストの投稿時刻を用いることで、疑似的に睡眠時間を推定することを考える。また、ユーザーの属性により体重や睡眠時間等の傾向が変わることも考慮し、ユーザーのプロフィール情報の言語分析についても取り組む。その他、生活習慣に関連するトピックとして、ユーザーの趣味に関する情報を分析する研究についても取り組む。

「食生活」については、SNS 上に投稿された食事画像を、言語表現と結びつけるための手法について研究する。そのために、Twitter API を用いて、画像付きツイートを、様々な食事カテゴリについて収集した「Twitter 食事画像データセット」を構築する。

健康状態に関するユーザーの感情状態を分析するための手法についても取り組む。闘病に関するブログデータを収集し、その中で、感情に関する記述を抽出するための手法の構築にも取り組む。

また、数値から抽出された言語表現を、実際に文書の検索に役立てることを目指すが、このさい、文書の中で重要な部分を抜き出すことを目指し、文書の文書構造（レイアウト）を正確に解析するための研究にも取り組む。

４．研究成果

（１）ユーザーの睡眠時間と投稿の関係分析について研究を進めた。手法としては、ユーザーの起床時のツイートを「おはよう」等の文字列を用いて特定し、その直前のツイートを就寝時のツイートと仮定することで、疑似的な睡眠時間として定義した。この疑似的な睡眠時間に対し、「8 時間以上」「8 時間未満」の 2 種類に分類し、各ツイートのラベル付けを行った。1,500 ツイートを訓練データ、100 ツイートをテストデータとして、SVM を用いて分類したところ、動詞のみを特徴量として用いることで、83%という高い分類精度となった。しかしながら、この結果を詳細に分析したところ、収集したツイートに bot が多く含まれており、実際に、負例の特徴語を調査したところ、「副業」「営業」「ニュース」「DM」等、bot に特徴的な語が多く含まれていた。これらのノイズが分類を容易にしてしまっていると考え、bot を含まないデータセットを新たに構築

した。具体的には、ツイートのユーザープロフィール欄を用い、学生のツイートのみを収集することを行い、また、人手により、自動的に抽出された起床ツイートが、本当に起床時のツイートか否かのチェックを行い、「もう少し寝る」等の、就寝中のツイート等のノイズを除去した。

結果として得られた、よりクリーンなデータセットに対し、訓練データ 1,817 ツイート、テストデータ 484 ツイートを用い、同様に SVM による分類を行った。

表 1 ・SVM による分類結果

使用特徴量	名詞	動詞	名詞＋動詞
分類精度	0.623	0.617	0.633

表 1 に示す通り、6 割以上の正解率を得られることを確認した。また、分類の重みの高い特徴量を確認したところ、「時間」「2 時」「時計」といった、夜遅い時刻であることにユーザーが気づくことを示す単語や、「テスト」「寝たい」といった、寝たいのに寝られない状況を示す単語、「寝落ち」等、きちんとした睡眠を取れていないことを示唆する単語等が得られた。

（2）ユーザーの食事習慣の分析のため、食事に関する言語と画像の関連分析用データセットを新たに構築した。具体的には、Twitter に投稿された画像を、投稿テキストと紐づけたデータを収集した。

既存の食事画像データセット UEC Food-256 を参考に、各カテゴリに対応する日本語を、同義語も含めて設定した。設定したカテゴリ語を Twitter API を用いて検索し、画像付きツイートを収集した。このさい、得られたデータには、食事と関係のない画像も含まれているため、画像認識で用いられる基本的なニューラルネットワーク VGG-16 及び、既存のデータセットから選択した画像データ（UEC Food 256 等の既存の食事画像データセット、および、ImageNet のうち、food カテゴリに属さない画像）を正例・負例として用いて独自に訓練したニューラルネットワークを用い、非食事画像のフィルタリングを行った。そのうち、検索結果数上位 120 のカテゴリについて、それぞれ 200 個の画像付きツイートを収集した。上位 10 個のカテゴリは、1 位から順に、ご飯／ラーメン／ピザ／寿司／うどん／ステーキ／餃子／納豆／パフェ／サンドイッチ、となった。

得られたデータには、例えば、「うどん」の検索結果画像に「うどんソフトクリーム」が含まれている、「アップルパイ」の検索結果画像に「アップルパイ味のお菓子」が含まれている等のノイズが含まれているため、これらのノイズを除去する手法について研究した。具体的には、それぞれの画像を VGG-16 を用いてベクトル化し、各カテゴリについて、全画像のベクトル平均を計算し、最も値の高い次元を求め、各カテゴリの代表次元とした。代表次元の値が最大値になる画像のみを残すことで、ノイズとなる画像の割合を減らせることを確認した。また、共起する画像の特徴量との関連度が高い文字列を収集したところ、「定食」「○○チップス」「○○お弁当人気投票」「自炊」「○○レシピ」といった、本来の料理カテゴリには無いが、特徴的な画像と関連する文字列を抽出できることを確認した。これを用いることにより、ノイズ除去に役立てることができると考えられる。

（3）数値と言語の関係を分析するための、複数の手法について検討・実装を行った。まず、ノンパラメトリックベイズモデルを応用し、数値と言語の統合的生成モデルを考案し、実装を行った結果、数値範囲に特徴的な言語表現の自動抽出や、言語表現に特徴的な数値表現の抽出等が行えることを確認した。また、同手法の地理情報と言語表現の関連抽出への適用も試みた。

その他、単語分散表現と数値の分布表現を統合する手法についても研究を進めた。学習された単語分散表現を文字列抽出に応用することにより、数値に特有の文字列表現を分散表現の観点から取得する手法について研究を行った。また、単純な連結による統合のみならず、単語分散表現の次元の中に数値の分布を埋め込むという新たな手法を開発した。これにより、複数の単位に関わる分布表現を固定次元の表現に埋め込むことが可能になった。また、得られた単語分散表現を用いることにより、関連する単語の類義語抽出において、同義語と対義語の弁別能力をある程度向上させることができることを確認した。

実際に、Twitter 上の文を「血圧」で検索した結果に、上記手法のうち、自動数値範囲推定及び各数値範囲の特徴表現抽出を行ったところ、例えば血圧の高い数値範囲に対し、「#を超え」「#オーバー」といった特徴表現を抽出することができ、また、これらの表現を用いて文書検索することで、実際に、対象の数値範囲に適した文書を検索できることを確認した。

（4）ユーザーが自らの闘病について記述した「闘病ブログ」から、特に糖尿病患者のブログに着目し、テキストを収集した。「時間」「食事」「運動」「薬」「血糖値」「HbA1c」「理由」「感情」の 8 つのカテゴリを設定し、ブログ中の、これらのカテゴリに属するキーワード及びキーフレーズにタグ付けすることにより、「闘病ブログコーパス」を作成した。単語分散表現や Sentence BERT をもとにしてキーワードのカテゴリ分類器を作成したところ、「理由」タグ（F 値 0.42）以

外のタグについて、0.65 以上の F 値を達成することを確認した。

（5）このほか、ユーザーの音楽に関する趣味を分析する手法として、ユーザーのプロフィール情報を言語的特徴量として用いることで、音楽アーティストの分散表現を学習する手法を開発し、アーティストの傾向分析へ応用できることを確認した。また、ユーザーが音楽番組を視聴した際の感想ツイートや、音楽共有サイトに投稿されたユーザーの感想テキストを用いてアーティストを分類する手法や、テレビ番組を視聴する際のユーザーの反応分析についても研究を進めた。

その他、ユーザーの趣味に関する研究として、「オンラインゲームのプレイヤー募集ツイート「飲食店へのレビューツイート」「観光地でのツイート」を分析する研究を行ったほか、「小説の自動要約」「音楽視聴者への推薦」といった研究も行った。

（6）ユーザーのプロフィール情報と投稿との関連を分析するために、プロフィール中の単語と投稿中の単語分散表現を同時に学習する手法を開発し、プロフィールを用いたツイート予測精度を通じて性能を検証した。

文書中の見出し語どうしの階層構造を判定するために、Wikipedia の見出し語についての分散表現を集計し、与えられた単語ペアについて、どちらが見出し語としてより上位に来るかを推定するアルゴリズムについて研究を行った。

テキスト解析の基礎技術として、アスキーアートの分類についても研究を進めた。Twitter やブログといった文書にはこれらの文字による視覚情報が多く含まれており、特にノイズ除去の前処理に効果を発揮することが期待できる。

５．主な発表論文等

〔雑誌論文〕　計3件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　2件）

１．著者名 Minoru Yoshida, Takumi Kojima, Kazuyuki Matsumoto, and Kenji Kita	４．巻 12(1)
２．論文標題 Toward Analyzing Relations between Sleeping Time and Social Networking Service Texts: Prediction of the Tweet Time Span Using the Last Tweet of the Day	５．発行年 2021年
３．雑誌名 International Journal of Advanced Intelligence	６．最初と最後の頁 1-9
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Kazuyuki Matsumoto, Seiji Tsuchiya, Takumi Kojima, Hiroya Kondo, Minoru Yoshida and Kenji Kita	４．巻 10
２．論文標題 Classification of Smartphone Application Reviews Using Small Corpus Based on Bidirectional LSTM Transformer	５．発行年 2019年
３．雑誌名 International Journal of Machine Learning and Computing	６．最初と最後の頁 148-157
掲載論文のDOI（デジタルオブジェクト識別子） 10.18178/ijmlc.2020.10.1.912	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Kazuyuki Matsumoto, Akira Fujisawa, Minoru Yoshida	４．巻 13
２．論文標題 ASCII Art Classification based on Deep Neural Networks Using Image Feature of Characters	５．発行年 2018年
３．雑誌名 Journal of Software	６．最初と最後の頁 559～572
掲載論文のDOI（デジタルオブジェクト識別子） 10.17706/jsw.13.10.559-572	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計19件（うち招待講演　0件／うち国際学会　12件）

１．発表者名 Minoru Yoshida, Shogo Kohno, Kazuyuki Matsumoto, Kenji Kita
２．発表標題 Visualization of the Artist Relations Using Twitter User Profiles
３．学会等名 the 6th International Conference on Fuzzy Systems and Data Mining (FSDM 2020)（国際学会）
４．発表年 2020年

１．発表者名
Fujisawa Akira, Kazuyuki Matsumoto, Kazuki Ohta, Minoru Yoshida and Kenji Kita

２．発表標題
ASCII Art Classification Model by Transfer Learning and Data Augmentation

３．学会等名
the 6th International Conference on Fuzzy Systems and Data Mining (FSDM 2020)（国際学会）

４．発表年
2020年

１．発表者名
Kirihara Taiga, Kazuyuki Matsumoto, Minoru Yoshida and Kenji Kita

２．発表標題
Keyword Extraction from TV Program Viewers Tweet Based on Neural Embedding Model

３．学会等名
the 6th International Conference on Fuzzy Systems and Data Mining (FSDM 2020)（国際学会）

４．発表年
2020年

１．発表者名
Ryu Mopuaa, Kazuyuki Matsumoto, Minoru Yoshida and Kenji Kita

２．発表標題
Construction of Annotated TOBYO Blog Corpus for Lifestyle Disease Analysis of Diabetic Patient

３．学会等名
The 15th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE20)（国際学会）

４．発表年
2020年

１．発表者名
Cai Xiaohan, 吉田稔, 松本和幸, 北研二

２．発表標題
Facts analysis of food tweets

３．学会等名
情報処理学会 第83回全国大会

４．発表年
2021年

１．発表者名
喜島 涼太，松本 和幸，吉田 稔，北 研二

２．発表標題
MBTI性格推定モデルの構築における感性情報の有効性

３．学会等名
日本感性工学会春季大会(JSAEE2021)

４．発表年
2021年

１．発表者名
Kazuyuki Matsumoto, Manabu Sasayama, Minoru Yoshida and Kenji Kita

２．発表標題
Emotional State Estimation by Dialogue History and Sentence Distributed Representation

３．学会等名
6th IEEE International Conference on Cloud Computing and Intelligence Systems(CCIS2019)（国際学会）

４．発表年
2019年

１．発表者名
Akira Fujisawa, Kazuyuki Matsumoto, Minoru Yoshida and Kenji Kita

２．発表標題
An Approach for Conversion of Japanese Emoticons into Emoji Based on Character-Level Neural Autoencoder

３．学会等名
Frontiers in Artificial Intelligence and Applications（国際学会）

４．発表年
2019年

１．発表者名
Kazuyuki Matsumoto, Mopaa Ryu, Minoru Yoshida and Kenji Kita

２．発表標題
Emotion Analysis on Weblog of Lifestyle Diseases

３．学会等名
the 2019 International Symposium on Signal Processing Systems (SSPS2019)（国際学会）

４．発表年
2019年

１．発表者名
Kazuyuki Matsumoto, Yuta Hada, Minoru Yoshida and Kenji Kita

２．発表標題
Analysis of Reply-Tweets for Buzz Tweet Detection

３．学会等名
Asia Pacific Society for Computing and Information Technology 2019 Annual Meeting (APSCIT 2019 Annual Meeting)（国際学会）

４．発表年
2019年

１．発表者名
松本 和幸，篠山 学，寺園 嶺，吉田 稔，北 研二

２．発表標題
インタビュー対話コーパスにおける発話の意図および感性の分析

３．学会等名
日本感性工学会春季大会

４．発表年
2020年

１．発表者名
高野翔吾，北　研二，吉田　稔，松本和幸

２．発表標題
文書分散表現を用いた音楽アーティストの特徴分類に関する研究

３．学会等名
情報処理学会第82回全国大会

４．発表年
2020年

１．発表者名
児嶋 拓己，吉田 稔，松本 和幸，北 研二

２．発表標題
ツイート内の発言による夜間活動休止時間の推定に関する研究

３．学会等名
情報処理学会第82回全国大会

４．発表年
2020年

１．発表者名
Minoru Yoshida, Kazuyuki Matsumoto and Kenji Kita

２．発表標題
Modeling Relations Between Profiles and Texts

３．学会等名
The 14th the Asia Information Retrieval Societies Conference (AIRS 2018)（国際学会）

４．発表年
2018年

１．発表者名
Akira Fujisawa, Kazuyuki Matsumoto, Kazuki Ohta, Minoru Yoshida and Kenji Kita

２．発表標題
ASCII Art Category Classification based on Deep Convolutional Neural Networks

３．学会等名
The 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)（国際学会）

４．発表年
2018年

１．発表者名
Kazuyuki Matsumoto, Minoru Yoshida and Kenji Kita

２．発表標題
Classification of Emoji Categories from Tweet Based on Deep Neural Networks

３．学会等名
The 2nd International Conference on Natural Language Processing and Information Retrieval (NLPIR2018)（国際学会）

４．発表年
2018年

１．発表者名
Horikawa Haruki, Minoru Yoshida, Kenji Kita, Kazuyuki Matsumoto

２．発表標題
Extended Dataset for Food Image Classification Based on Twitter

３．学会等名
Hong Kong International Conference on Engineering and Applied Science (HKICEAS 2018)（国際学会）

４．発表年
2018年

１．発表者名
藤野 尚也，松本 和幸，吉田 稔，北 研二

２．発表標題
Word Mover's Distanceを用いたコーパス拡張による感情推定精度向上の検討

３．学会等名
言語処理学会第25回年次大会(NLP2019)

４．発表年
2019年

１．発表者名
羽田 優太，松本 和幸，吉田 稔，北 研二

２．発表標題
リプライを用いたバズツイートの分類

３．学会等名
言語処理学会第25回年次大会(NLP2019)

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	北　研二 (KITA Kenji) (10243734)	徳島大学・大学院社会産業理工学研究部（理工学域）・教授 (16101)	
研究分担者	松本　和幸 (MATSUMOTO Kazuyuki) (90509754)	徳島大学・大学院社会産業理工学研究部（理工学域）・准教授 (16101)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関