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　４ 年　 ６ 月 １３ 日現在

機関番号：３２６８９
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１９８３

研究課題名（和文）ソーシャルメディアの時間的・意味的分析を活用した知識グラフの構造的拡張

研究課題名（英文）Structural extension of knolwedge graph utilizing temporal and semantic analysis of social media

研究代表者
岩井原　瑞穂（Iwaihara, Mizuho）

早稲田大学・理工学術院（情報生産システム研究科・センター）・教授

研究者番号：４０２５３５３８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：　知識蓄積型ソーシャルメディアであるWikipediaからは，計算機利用が容易な構造的データが知識グラフとして抽出され，検索結果の分類や種々の自然言語処理に活用されている．知識グラフを充実させるためのWikipedia記事のマイニングにおいて，リンクやリストなどの構造情報を活用し，さらに拡張する手法が必要である．
本研究では，Wikipediaにおいて，併合すべき記事対の予測および新たなリンクを予測する手法を開発した．テキストからのキーフレーズ抽出について，訓練済み言語モデルを用いた手法を開発し従来を上回る精度を示した．センチメント分析を応用したツィートの著者推定を行う手法を開発した．

研究成果の学術的意義や社会的意義
ウェブからの有用な情報の抽出は，日々生成される膨大なデータを整理分類する基礎的段階を含む．テキスト分類は伝統的に多くの手法が提案されてきたが，新たな形態のテキストとして，Wikipediaの記事の階層的構造や，ツィートのハッシュタグ，さらにこれらの時系列的要素などの課題が出現している．一方，訓練済み学習モデルと呼ばれる深層学習を元にした手法が，従来手法を一変させつつある．本研究では，キーフレーズ抽出，リンク予測，階層的分類等の問題および知識グラフの応用について幅広く研究を行い，いくつかの問題では従来を上回る性能を示すことができた．

研究成果の概要（英文）：Wikipedia is known as the largest social media collecting knowledge, from which knowledge graphs are extracted as computer-readable structured knowledge models. Knowledge graphs are utilized for search result enrichment and various natural language tasks. For developing high-quality knowledge graphs from Wikipedia, structured data such as lists and categories need to be utilized.
In this research, we developed new methods for predicting Wikipedia article pairs that should be merged, and pairs that should have links. For extracting keyphrases from article texts, we developed a method utilizing pretrained language models, improving known records on this task. We also proposed new methods for authorship attribution on tweets, utilizing text sentiment.

研究分野：メディア情報学

キーワード：データマイニング　テキストマイニング　情報抽出　知識グラフ　時系列分析

様　式　Ｆ－１９－２

１．研究開始当初の背景

代表的な知識蓄積型ソーシャルメディアである Wikipedia は，ユーザの投稿に基づく巨大な百科事典であるとともに，計算機利用が容易な構造的データが知識グラフとして抽出され，検索結果の分類や種々の自然言語処理に活用されている．

Wikipedia はリンク関係やカテゴリ，infobox など計算機利用が容易な構造的データを含んでいるため，記事の項目をノードとし項目間の関連を枝とする知識グラフが抽出できる．知識グラフは検索結果の分類や問い合わせの補完や，自然言語処理の多様なタスク等に広く活用されており，代表的な知識グラフとして WikiData や Google Graph がある．また，Wikipedia の各記事は過去のバージョンが編集履歴として蓄積・公開されているため，時系列データとしても利用できる．

一方 Twitter や Facebook などの交流型ソーシャルメディアからは，共感，興味など感情的反応を含めた反響の大きさを測定でき，Wikipedia 記事の編集履歴と対比させることにより，記事の重要性・公平性や充実度の評価が可能であり，知識グラフの充実化に活用することができる．ツィートでは，短文でくだけた表現や絵文字の多用といった特徴があり，またハッシュタグを自由に付加することにより，同じハッシュタグを含むツィートが連なって表示され分類タグとして機能し，同じトピックのツィートを効率よく検出できる．ハッシュタグおよびカテゴリラベルは，ユーザが自由に生成でき，またキーフレーズ（特徴的な語句）と分類ラベルという両方の機能を持つという特長がある．

２．研究の目的

本研究はソーシャルメディアのコンテンツの時系列分析・意味的類似度分析を統合した新たな知識構造抽出技術の開発と，その応用展開を目的とする．知識グラフを充実させるためには，Wikipedia 記事のマイニングにおいて，リンクやリスト，カテゴリなどの構造情報を活用し，さらに拡張する新たな手法が必要である．

本研究では，(1)知識グラフの構造的拡張において，(1-a) 記事間のリンク予測および記事の分離統合予測問題，(1-b) Wikipedia リストの要素帰属問題およびテーブルスキーマ生成問題，(1-c) 実体リンキングへの応用からなる新たな課題に取り組む．また記事の時系列変化に着目した，(2) 編集履歴からの特徴的語句の抽出に取り組む．(3)ソーシャルメディアにおけるセンチメントの集約表現では，ツイートなどのソーシャルメディアにおける関心やムードの簡潔な集約表現を開発し，Wikipedia の成長過程との対比を可能にすることを目指す．

３．研究の方法

自然言語処理分野では，BERT に代表される，巨大なコーパスで深層学習モデルを事前に訓練した学習済み言語モデルが開発されて以来，比較的少量な訓練データで目標タスクに finetuning する方式が発展し，従来の数々の記録を塗り替えている．学習済み言語モデルは，自然言語文からの関連の抽出，カテゴリへの文書分類，文書からのキーフレーズの抽出など知識グラフに関連したタスクにも応用されている．

一方，グラフデータにおいても，グラフノードの近傍関係を，連続的なベクトル空間に埋め込むことにより，グラフノードの近接性や構造的類似性を求める手法や，リンク予測手法が開発されている．本研究では，これらの手法を組み合わせることにより，知識グラフの構築と拡張のための学習済み言語モデルを活用した新たな手法の開発を目的とする．

４．研究成果

(1) Wikipedia では，多くの記事が並行して入力されるため，記事の内容の重複が頻繁に生じている．そのため編集者間で議論し，記事の整理統合を随時行っている．記事の分離統合予測問題として，自動的に重複した内容の記事を発見し，ひとつの記事に統合すべきかあるいは独自の内容があるため分離した記事のままにするかを判定する問題を新たに定義し，その計算手法を開発した．本問題は単純な語句の重なりや類似度計算では判定できないため，文からパラグラフまで，様々なサイズの文章ペアの類似度分布を入力とする畳み込み学習による予測する手法を開発した．長大な記事における，節単位の類似度の分布から得られる特徴量，および記事のトピックに適応させた単語埋め込みを用いて，弁別器を訓練することにより，標準的な手法よりも精度良く予測できることを示した．訓練データとして，実際の分離統合の履歴を用いている．

(2) 知識グラフにおけるリンクの補完問題について，実体の属するクラス集合を新たに利用することにより，精度を向上できることを示した．記事間のリンク構造を，時刻印に従ったランダムウォークを行うことにより，記事のグラフ構造の埋め込みベクトルを学習し，それを用いてリンク予測を行う手法を提案した．従来のランダムウォークに対し，記事の意味的類似度も考慮することにより，精度が向上できることを示した．

(3) キーフレーズ抽出に関しては，キーフレーズ候補の共起関係のグラフに PageRank アルゴリズムを適用する TextRank や，単語埋め込みベクトルに同手法を適用する EmbedRank が知られている．Wikipedia の記事は時間とともに追加・編集されるため，バースト的に編集され

るキーフレーズを抽出すれば，キーフレーズの時間的変化やトピックの発展を捕捉することができる．本研究では Wikipedia の記事集合の履歴から，時間的に変化するキーフレーズを抽出する方法を開発している．一方，テキストの要約はキーフレーズ抽出・生成と関連の深いタスクであり，これまでに Wikipedia の節ごとの内容を要約する節タイトルを，深層学習モデルにより生成する手法を提案している.

(4) 文書を特徴づけるキーフレーズは，文書分類の重要な要素であるとともに，新たなカテゴリラベルやハッシュタグとして用いることができるため，知識グラフの構造的拡張に有用である．キーフレーズ抽出に関し，対象文書とキーフレーズ候補のペアを順位付けする問題として BERT に学習させる手法を考案した．さらに，長い文書を対象とするときに，キーフレーズを含む可能性が高い文をあらかじめ選択して文書の要約を生成し，その要約にキーフレーズ抽出を適用する手法を示した．要約のためにもう一つ別の BERT モデルを訓練する．この事前要約と順位付けの組み合わせにより，TextRank および EmbedRank を大きく上回る精度でキーフレーズ抽出することができた(ICALD21 で発表).

(5) 知識グラフにあるべき枝が欠けている場合に補完する，知識グラフのリンク予測が知られている．これには，有向グラフのノードの近傍関係を保存したまま低次元ベクトル空間の点集合に写像する，グラフ埋め込み手法がある．この手法は友達関係など時間とともに発展するソーシャルグラフなど時系列グラフにおいて，将来の枝を予測するタスクにも用いられている．既存の手法では主にグラフの構造的情報と，ノードの属性値を対象としており，Wikipedia 記事や，ツイート，論文などの比較的長いテキスト情報を持つ場合のリンク予測手法は知られていなかった．本研究では，Wikipedia 記事の間でのリンクを予測する問題において，記事のタイトルおよびカテゴリラベルのテキストを抽出して，記事のペア間のリンクの有無と，テキストの意味的類似性との関連性を学習済み言語モデルで学習させた．これにより，グラフ構造のみによる埋め込み手法よりも，リンクの予測精度を向上させた(KDIR21 で発表).

(6) マルチタスク学習という，１つの BERT モデルを複数の関連したタスクで共有して訓練し，また学習パラメータを学習過程で変化させることにより，テキストが主観的か客観的かに分類する主観性判別タスクにおいて，最高の精度を達成することができた(APWeb20).

(7) ツィートの著者推定とは，与えられたツィートがどのユーザによるかを求めるタスクであり，重複アカウント検出やスパム検出等の応用がある．これまで，CNN による深層学習モデルをもとに，正負のセンチメント極性分布やテキスト長など著者の書くスタイルに対応する特徴量を定義したモデルを提案している(CBDCom 19,APWeb20)．さらにツィートの表記のゆらぎにも頑強になるように finetuning を行った学習済み言語モデルを利用することにより，ツィートの著者推定ではこれまでの最高の精度を達成している(APWeb21).

(8) 文書をカテゴリに分類してカテゴリラベルを与える文書分類は，知識グラフ構築および拡張の基本的なタスクである．複数のカテゴリのうちいずれかに分類する多クラス分類に対し，多ラベル分類では１つの文書が複数のラベルを持つことが可能であり，ラベルごとの有無を判定する二値分類器の集合を構築することになる．多ラベル分類では，ラベルの空間が非常に大きくなり，一般に多クラス分類よりも難しい問題であるが，Wikipedia のカテゴリラベルやツィートのハッシュタグ，科学技術論文に見られるように，多ラベル分類がより実際的である．一方，訓練のために，文書集合に正解ラベルを与える作業はコストがかかり，さらに新たに加えられた文書に追随しにくい等の問題点がある．少数のラベル付けされた訓練文書集合と，多数のラベルのない文書集合が存在するのが現実的な状況である．従来から，弱教師付き学習として，少数のラベル付き文書をもとに，ラベル無し文書の正解を疑似ラベルとして推定し，確信度の高い疑似ラベルを用いて分類器を自己訓練するアプローチがある．文書分類では，類似した文書は類似したラベルを持つという仮定のもと，テキストの類似度によりラベルを伝搬させてゆく LPA や TRAM という手法が知られている．一方，文書間の意味的類似度は学習済み言語モデルにより求めることができるが，目標のラベルに適応するよう類似度を finetuning するアプローチが有効である．そのため本研究ではラベル伝搬で得られた疑似ラベルを BERT の finetuning の訓練に利用して意味的空間と分類器を改良し，そこで得られた疑似ラベルを次のステップで伝搬させるという，意味的空間の finetuning とラベル伝搬を交互に繰り返す手法を提案し，従来手法を大幅に改良する結果を得ている（ICADL21 で発表）.

(9) Wikipedia やウェブページには，テーブル形式による情報表現が多く用いられている．テーブルからの情報抽出には，テーブルの各行や列(属性)が意味するものを推定する必要がある．属性の並びからデータ型が文字列であることは容易にわかるが，文字列が人名を表すのか，地名を表すのか，さらに特定の組織の構成員を表すのか，という意味的型付けの問題が存在する．本研究では，テーブルの属性値として現れた名前を元に，正確な意味的型付けを行うために，

知識グラフを検索して属性値の表す実体を求める手法を開発した．さらに意味的型付けの弁別器を訓練するには，テーブルの属性値のみでは不足するため，知識グラフから得られた，同じ型を持つ実体を加えるデータ拡張により，深層弁別器の訓練が可能であることを示した．さらに，単一の属性のみだけではなく，同じテーブル内の他の属性値として現れている実体名も用い，テーブルをひとつのトークン列にエンコードして，BERT で学習させる手法を新たに開発し，従来手法よりも意味的型付けの精度を向上できることを示した．

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Renzhi Wang, Mizuho Iwaihara	４．巻 28
２．論文標題 Detection of Mergeable Wikipedia Articles Utilizing Multiple Similarity Measures	５．発行年 2020年
３．雑誌名 Journal of Information Processing	６．最初と最後の頁 178-191
掲載論文のDOI（デジタルオブジェクト識別子） 10.2197/ipsjjip.28.178	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14件（うち招待講演　0件／うち国際学会　10件）

１．発表者名 Tingyi Liu and Mizuho Iwaihara
２．発表標題 Supervised Learning of Keyphrase Extraction Utilizing Prior Summarization
３．学会等名 Proc. 23rd Int. Conf. Asia-Pacific Digital Libraries (ICADL21),online, LNCS Vol.13133, pp. 157-166, Dec. 2021.（国際学会）
４．発表年 2021年

１．発表者名 Zhewei Xu and Mizuho Iwaihara
２．発表標題 Integrating Semantic Space Finetuning and Self-training for Semi-supervised Multi-label Text Classification
３．学会等名 Proc. 23rd Int. Conf. Asia-Pacific Digital Libraries (ICADL21),online, LNCS Vol.13133, pp. 249-263, Dec. 2021.（国際学会）
４．発表年 2021年

１．発表者名 Jiaji Ma, and Mizuho Iwaihara
２．発表標題 Link Prediction for Wikipedia Articles based on Temporal Article Embedding
３．学会等名 Proc. 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR,pp. 87-94, Oct 2021.（国際学会）
４．発表年 2021年

１．発表者名
Xiangyu Wang and Mizuho Iwaihara

２．発表標題
"Integrating RoBERTa Fine-Tuning and User Writing Styles for Authorship Attribution of Short Texts

３．学会等名
Proc. 5th APWeb-WAIM Joint Conference on Web and Big Data (APWeb-WAIM 2021), LNCS 12858, pp. 413-421, Aug. 2021.（国際学会）

４．発表年
2021年

１．発表者名
A Diya, Mizuho Iwaihara

２．発表標題
Keyphrase Generation by Utilizing BART Finetuning and BERT-Based Ranking

３．学会等名
DEIM Forum G24-3 , Online, March 2022

４．発表年
2022年

１．発表者名
Qin Jiaxin, Mizuho Iwaihara

２．発表標題
Annotating Column Type Utilizing BERT and Knowledge Graph Over Wikipedia Categories and Lists

３．学会等名
DEIM Forum G33-1, Online, March 2022.

４．発表年
2022年

１．発表者名
Huang Zeping, Mizuho Iwaihara

２．発表標題
Authorship Attribution Based on Pre-Trained Language Model and Capsule Network

３．学会等名
DEIM Forum H33-4, Online, March 2022.

４．発表年
2022年

１．発表者名
Li Peining, Mizuho Iwaihara

２．発表標題
Column Type Detection Based on Pretrained Language Models with Various Column Encodings

３．学会等名
DEIM Forum E43-3, Online, March 2022

４．発表年
2022年

１．発表者名
Wenjing Huang, Rui Su and Mizuho Iwaihara

２．発表標題
Contribution of Improved Character Embedding and Latent Posting Styles to Authorship Attribution of Short Texts

３．学会等名
4th APWeb-WAIM Joint Conference on Web and Big Data (APWeb-WAIM 2020) (国際学会)

４．発表年
2020年

１．発表者名
Hairong Huo and Mizuho Iwaihara

２．発表標題
Utilizing BERT Pretrained Models with Various Fine-tune Methods for Subjectivity Detection

３．学会等名
4th APWeb-WAIM Joint Conference on Web and Big Data (APWeb-WAIM 2020) (国際学会)

４．発表年
2020年

１．発表者名
Zihang Chen, Mizuho Iwaihara

２．発表標題
Detection of Editing Bursts and Extraction of Significant Keyphrases from Wikipedia Edit History, In Big Data Analyses, Services, and Smart Data, Advances in Intelligent Systems and Computing book series

３．学会等名
Advances in Intelligent Systems and Computing book series (国際学会)

４．発表年
2020年

１．発表者名
Xingyu Chen, Mizuho Iwaihara

２．発表標題
Weakly-Supervised Neural Categorization of Wikipedia Articles

３．学会等名
Proc. ICADL2019, LNCS11853, pp. 16-22, Nov. 2019.（国際学会）

４．発表年
2019年

１．発表者名
Junhao Li and Mizuho Iwaihara

２．発表標題
Two-Encoder Pointer-Generator Network for Summarizing Segments of Long Articles

３．学会等名
The Asia Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conf. Web and Big Data (APWeb-WAIM 2019), LNCS 11641, pp. 299-313, Chengdu（国際学会）

４．発表年
2019年

１．発表者名
Patamawadee Leepaisomboon, Mizuho Iwaihara

２．発表標題
Utilizing Latent Posting Style for Authorship Attribution on Short Texts

３．学会等名
Proc. IEEE Int. Conf. Cloud and Big Data Computing (CBDCom 2019), pp.1015-1022, Fukuoka（国際学会）

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

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