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　 ４ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2018～2020
課題番号：１８Ｋ１１９８２

研究課題名（和文）体験談アーカイビングにおける地理的位置への言及に基づいた自動索引付けに関する研究

研究課題名（英文）A Study on Automatic Indexing Based on Textual Mentions to Geographical Location in Story Archiving

研究代表者
乾　孝司（INUI, Takashi）

筑波大学・システム情報系・准教授

研究者番号：６０３９７０３１

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）：本研究課題は、文書コンテンツ内で言及されている地理的位置を索引付けし、地理的位置による文書検索技術を開発することを目的とする。主な研究成果として以下が挙げられる。（１）文書内の単語情報と文書内単語に対応した画像情報を併用することで、未知語事例に強い深層学習ベース地名抽出モデルを開発した。（２）住所階層性に着目したデータ拡張を施した単語分布に基づく地名の地理的位置の特定（地名の曖昧性解消）モデルを開発した。（３）上記モデルを統合することで、文書コンテンツ内で言及されている地理的位置の索引付けに必要な地理的位置情報をある程度の性能で自動特定する技術を実現した。

研究成果の学術的意義や社会的意義
本研究課題は、大規模自然災害アーカイブにおいて、従来技術では地理情報システムとの親和性の低かった文書コンテンツに対して、特定の被災地域に限定したコンテンツ検索を実現するための技術開発を目的としたものである。本研究課題で得られた成果を活用することにより、自然災害に対する防災・減災対策や、自然災害からの復旧・復興事業に資する情報へのアクセス効率が従来よりも向上することが期待される。

研究成果の概要（英文）：This research project aims to develop a document retrieval technology by geographic location by indexing geographic locations mentioned in the document contents. The main research results are as follows. (1) We developed a deep learning-based geographic name extraction model that is especially robust to unknown words by using word information in documents and image information corresponding to words. (2) We developed a model for identifying the real-world geographic location of place names (place name disambiguation) based on word distributions with data expansion focusing on address hierarchy. (3) By integrating the above models, we have developed a technology to automatically identify geographic location information required for indexing geographic locations mentioned in document contents with a certain level of performance.

研究分野：自然言語処理

キーワード：文書ジオロケーション　エンティティ・リンキング　地名抽出　固有表現抽出　Toponym resolution

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

巨大地震や大型台風といった同時期多発的に甚大な被害をもたらす大規模自然災害に関する記録資料は、復旧・復興事業や、防災・減災対策に資する貴重な国家的財産である。例えば、国立国会図書館のインターネット資料収集保存事業では東日本大震災アーカイブ（http://kn.ndl.go.jp/）として、被災地の当時の様子を綴った被災者の体験談データや被災地の様子を記録した写真データ類が 100 万件以上アーカイビングされている。このような大規模自然災害アーカイブにおいてはある特定の被災地域に限定してコンテンツを検索したいという利用者（復旧・復興事業，防災・減災対策の従事者）からの検索要求がある。これまで、メタ情報に基づいて特定の地理的位置に検索対象を絞り込むメタ情報検索がこの要求に間接的に応えてきたが、アーカイブされた体験談の中で言及されている被災地域が必ずしもメタ情報と一致する保証はなく、現在までのところ、アーカイブ利用者の検索要求に直接的に応える技術は確立されていないのが実情である。

２．研究の目的

本研究課題では、蓄積されたアーカイブ・コンテンツが死蔵されることなく有効活用される検索環境として、アーカイブ・コンテンツのうち体験談コンテンツに注目し、言及されている被災地域に限定して体験談を検索できる機能の実現を目標に、体験談データに対して被災地域（地理的位置）情報を高精度に自動索引付けする技術を開発することを目的とする。アーカイブ・コンテンツは多様なメディアで構成されている。これらのうち、画像／映像データについて、それらの多くは撮影機器がもつ GPS 機能などによって撮影時の場所情報を自動的かつ正確にメタ情報として付与でき、既に利用者の検索要求を十分に満たしていると言える。そのため本研究では、研究対象として画像／映像データではなく言語コンテンツである体験談データを対象とする。

３．研究の方法

体験談で言及されている被災地域を正確に特定し、その情報に従って体験談を索引付けする処理の概略を図１に示す。本研究では，この索引付け処理に不可欠となる２つの要素技術、地名抽出および地理的位置の特定（地名の曖昧性解消）、の開発をおこない、両技術に基づいた索引付け機能を実現することを目指す。各要素技術の詳細を以下に示す。

地名抽出とは、体験談テキストから地理的位置を示す地名を正確に自動抽出する技術である。体験談テキストにおいては、「茨城県水戸市笠原町に行きました」のように完全な住所表記が書かれるとは必ずしも限らず、「水戸に行きました」のように地名が略記されることがある。この時、テキスト中には地名と同じ名称であるが地理的位置を示さない言及も存在するため、住所データベースと照合するだけでは地名を正確に抽出できない。そのため、地理的位置を示す地名に関する言及のみを正確に自動抽出する技術が必要になる（例えば「私は水戸さんと一緒に…」の「水戸」は人名であり、抽出してはならない）。

地理的位置の特定（地名の曖昧性解消）は、抽出された地名があらわす正確な地理的位置を自動特定する技術である。都道府県名がユニークに命名される一方で、その内部地域である市区町村・大字（おおあざ）レベルの地名は同じ名称の地名が各所に多数存在する。そのため、地理的位置を正確に索引付けするには、抽出された地名が具体的にどの地理的位置をあらわすかを正確に自動特定する技術が必要になる（例えば「水戸」という名称の地域は図１の住所 DB に４箇所登録されており、索引付けするには正しい地域の特定が必要である）。

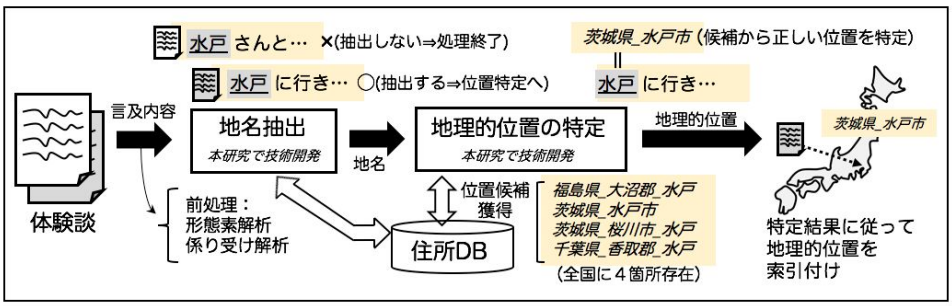

図１　処理の流れの概要

４．研究成果

まず、地名抽出の研究成果について述べる。我々はこれまでに、Wikipedia を Gazetteer 辞書として活用した条件付確率場を基づく固有表現抽出手法（以下、既存手法）を開発しており、これを地名抽出向けに改良する方針で開発を進めた。既存手法では、入力文書中の単語系列のうち、部分的な要素にしか辞書情報を反映できない問題があった。この問題を解決するために、本

研究ではカテゴリグラフカーネル（CGK）を利用した新しい辞書素性の構築手法を提案し、地名抽出に適用した。CGK とは、Wikipedia のエントリに対してあらかじめ指定した基底カテゴリへの所属確率を推定するアルゴリズムであり、我々は地名関連エントリを基底カテゴリに指定することで CGK を用いた。さらに、標準の CGK では地名以外の不要カテゴリの扱いが難しいため、新たに不要カテゴリの設定方法を提案した。提案手法の有効性を検証するため、拡張固有表現タグ付きコーパスを評価用データセットに用いた評価実験を実施した。実験結果を表１に示す。ここでは５種類の辞書利用の結果を示している。Wikipedia を Gazetteer 辞書とする「一般辞書」、一般辞書で Wikipedia の見出し曖昧性を考慮した「曖昧辞書」、一般辞書を地名に特化させた「地名辞書」はベースラインの辞書利用法であり、そして、従来の標準的 CGK に基づく「CGK-Location」、および CGK に対して不要カテゴリの扱いを改良した「CGK-Greedy」が提案手法である。表に示す通り、CGK を利用した提案手法は CGK を用いない既存手法よりも地名抽出性能が向上することを確認した。

表１　条件付確率場と辞書利用に基づく地名抽出の結果

利用辞書	再現率	適合率	F1 値
一般辞書	0.869	0.848	0.857
地名辞書	0.870	0.847	0.857
曖昧辞書	0.870	0.849	0.858**
CGK-Location	0.870	0.848	0.858
CGK-Greedy	0.869	**0.851**	**0.859****

つづいて、抽出手法の基本モデルを条件付確率場（CRF）から深層学習ベースのモデルへ変更した。具体的には固有表現抽出（地名抽出）のような系列データ処理に有効である LSTM ユニットを採用した双方向 LSTM-CRF（Bi-LSTM-CRF）に基づく地名抽出技術の開発をおこなった。 これにより、最終的な地名識別部は CRF で変更はないが、特徴抽出部が改良され、結果として、CRF に基づく手法よりも抽出性能の改善（F 値で 85.9→88.3）を達成した。

さらに、深層学習ベースのモデルは多様な特徴量を柔軟に組み込める利点を活かし、特徴量としてこれまでの言語情報に加えて、画像情報を加える方法を開発した。具体的には入力文書内の単語画像情報を言語情報と統合して利用する地名抽出の改良モデルを提案した（図２）。改良モデルでは、文書中の単語に対して、それをクエリとした画像検索を実行し、単語に対応する画像情報を取得する。そして、それを画像ベクトルへと変更の後、元の単語ベクトルと統合する。

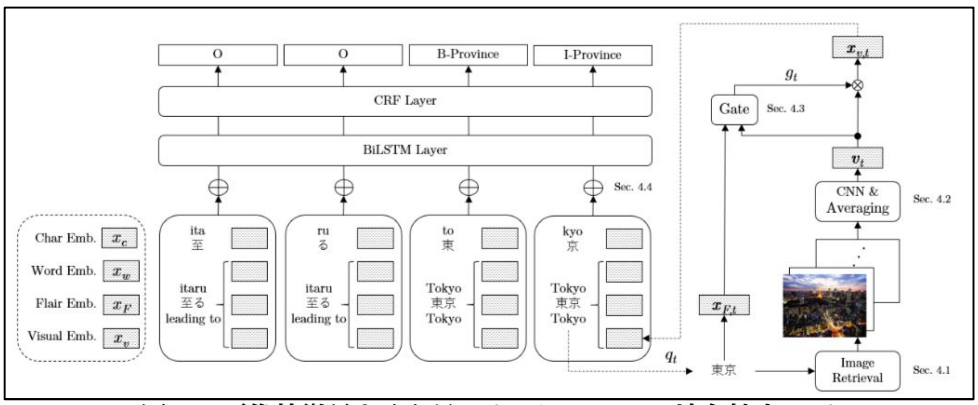

図２　画像特徴量を取り込んだ Bi-LSTM-CRF 地名抽出モデル

先述の実験と同じ評価用データを用いた評価実験を実施した。実験結果を表２に示す。表中の「Baseline」は画像情報を取り込まない Bi-LSTM-CRF の結果である。また、「Visual(Simple)」は画像情報と言語情報を単純に統合した場合の結果であり、「Visual(Gate)」はゲート機構を導入することで画像情報の取り込み量を学習した場合の結果である。評価実験の結果、提案手法は画像情報を取り込まない標準的な Bi-LSTM-CRF よりも高い F 値を達成した。また、特に未知事例が多く発生する City カテゴリの地名に注目し、評価データを既知事例と未知事例に分割して評価した結果、提案手法は特に未知事例に対して性能向上が顕著であることを確認した(表３)。

次に、地名の地理的位置特定（地名の曖昧性解消）の研究成果について述べる。これは、地名抽出によって抽出された地名があらわす正確な地理的位置を自動特定する技術であり、本研究では特に、同じ名称の地名が複数の地理的位置において使われている場合の曖昧性解消を取り扱う。図３に、City カテゴリの地名における曖昧候補数の分布を示す。図から、曖昧性があったとしても多くの場合は、２箇所あるいは３箇所の地理的位置間での衝突であることがわかるが、それ以上の候補数をもつ地名も少なからず存在することがわかる。

表２　画像情報を取り込んだ Bi-LSTM-CRL に基づく地名抽出の結果

Model	Prec.	Recall	F1
Baseline	87.33	89.47	88.38
Visual (Simple)	**90.20**	87.78	88.97*
Visual (Gate)	89.33	**90.01**	**89.67****

表３　未知事例に対する地名抽出の結果(City カテゴリ)

	Model	Precision	Recall	F1
Seen	Baseline	84.99	79.29	82.04
	Visual (Gate)	90.07 (+5.08)	80.05 (+0.76)	84.77 (+2.73)
Unseen	Baseline	68.54	54.95	61.0
	Visual (Gate)	75.61 (+7.07)	55.86 (+0.91)	64.25 (+3.25)

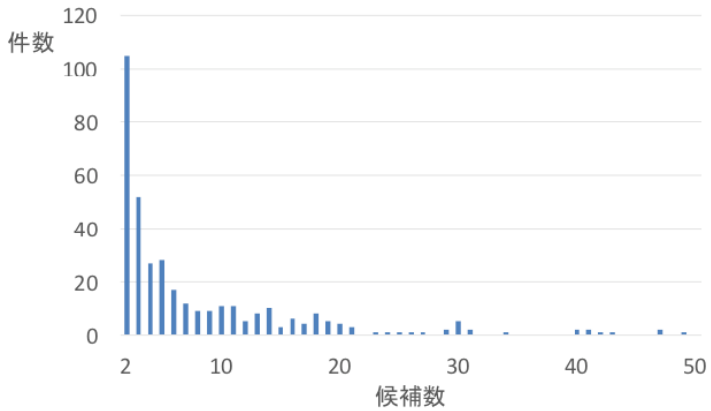

図３　City カテゴリの地名がもつ曖昧候補数

先行研究である POPULATION 法では人口データを利用するが、大字（おおあざ）などの細粒度地域に関しては高被覆な人口データの整備・管理が困難という問題があった。この問題を解決するために、人口データに替えて関連文書群中での地名の言及回数に基づく地理的位置特定手法（MENTION_COUNT 法）を開発した。これは、関連文書群において言及回数が多い地理的位置ほど特定対象となっている言及の地理的位置になりやすいという仮定に基づく。また、特定対象となっている言及の局所文脈は強い手がかりとなることが事前調査からわかっていることから局所文脈情報もあわせて利用する。さらに、地名の地理的位置特定課題は課題の性質上疎データ問題に陥りやすいという問題がある。そこで、既存手法である TRIPDL 法に対して、教師データと教師なしデータを併用できる改良手法を開発した。この手法では、文書中の都道府県レベルの言及地名情報に従って疑似教師データを構築し、通常の教師データと混合する。評価実験の結果を表４に示す。表中の「TRIPDL+」が提案手法であり、データ併用方法によって、常に併用するソフト設定と、教師データが不足する場合のみ併用するハード設定の結果を示している。この表から、提案手法は POPURATION 法よりも有効であることが確認でき、TRIPDL+ハード設定が最良の正解率を達成していることがわかる。

地理的位置で文書を索引付けする際、文書が短い場合は上述の要素技術に基づく手続きとは別に、直接、文書単位で地理的位置を特定することが有効な場合がある。そこで、深層学習に基づく文書単位での地理的位置特定手法の技術開発をおこなった。まず、既存手法である deepgeo 法の実装、追試をおこない、その問題点を精査した。その結果、人間であれば明らかな手がかりフレーズが文書に含まれている場合であってもその手がかりを見逃す誤りが散見していることを確認した。この問題を解決するために、図４のように、地名等のフレーズに対してインジケータ付与処理を前処理として施したのち deepgeo モデルを学習する手法を提案し、評価実験によりその有効性を検証した。その結果を表５に示す。インジケータ付与処理は辞書参照に基づく付与と文脈を考慮する形態素解析に基づく付与の２種類を試したが、どちらにおいてもインジケータを付与しない従来モデルよりも高い精度を達成することを確認した。

表４　地名の地理的位置特定の評価実験結果

手法	拡大正解率	正解率	正解	誤選択	出力なし	正解なし
RANDOM	24.4	22.2	384	1196	0	152
POPULATION	57.0	56.9	985	207	388	152
MENTION_COUNT	73.0	62.1	1,076	451	53	152
TRIPDL	69.5	69.1	1,197	88	295	152
ソフト設定						
TRIPDL+（$\alpha = 0$）	71.0	60.3	1,045	482	53	152
TRIPDL+（$\alpha = 0.2$）	46.6	36.6	634	898	48	152
TRIPDL+（$\alpha = 0.4$）	48.8	39.0	675	857	48	152
TRIPDL+（$\alpha = 0.6$）	51.8	43.1	747	785	48	152
TRIPDL+（$\alpha = 0.8$）	56.9	48.8	845	687	48	152
TRIPDL+（$\alpha = 0.9$）	61.7	54.4	943	589	48	152
TRIPDL+（$\alpha = 0.99$）	74.2	70.0	1,212	320	48	152
TRIPDL+（$\alpha = 0.999$）	78.2	75.1	1,300	232	48	152
TRIPDL+（$\alpha = 1.0$）	69.5	69.1	1,197	88	295	152
ハード設定						
TRIPDL+（$\alpha = 0/1$）	79.0	76.0	1,316	216	48	152

処理前　=　早くつきすぎて今治観光してる @ Imabari Castle

処理後　=　早くつきすぎて<loc>今治<loc>観光してる @ Imabari Castle

図４　インジケータ付与処理の例

表５　文書単位での地理的位置特定の実験結果

手法	分類精度
deepgeo 法	0.663
辞書インジケータ付 deepgeo 法	0.674
MeCab インジケータ付 deepgeo 法	**0.677**
人手 (参考上限値)	0.767

最後に、本研究課題では、文書が標準的な長さの場合は２つの要素技術、地名抽出および地理的位置の特定（地名の曖昧性解消）によって、また、文書が短い場合は直接文書単位で地理的位置を特定する技術を開発し、文書索引付けに必要な地理的位置情報をある程度の性能で自動特定する技術を実現できた。以上から、当初研究目標で述べた項目について概ね達成できたと言える。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計6件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Takuya Komada and Takashi Inui
２．発表標題 An Element-wise Visual-enhanced BiLSTM-CRF Model for Location Name Recognition
３．学会等名 The 3rd International Workshop on Spatial Language Understanding（国際学会）
４．発表年 2020年

１．発表者名 陰山宗一，駒田拓也，乾孝司
２．発表標題 ニューラル日本語固有表現認識における格フレームの有効性検証
３．学会等名 言語処理学会第27回年次大会
４．発表年 2021年

１．発表者名 関龍，乾孝司
２．発表標題 新聞記事中の地名に対する地理的位置推定における有効な素性の調査
３．学会等名 第33回人工知能学会全国大会
４．発表年 2019年

１．発表者名 平川冬尉，乾孝司
２．発表標題 日本語地理的位置推定課題におけるインジケータ付deepgeo法の提案と評価
３．学会等名 第34回人工知能学会全国大会
４．発表年 2020年

１．発表者名 関龍，乾孝司
２．発表標題 局所文脈と関連文書を用いた地名に対する地理的位置の同定
３．学会等名 第32回人工知能学会全国大会
４．発表年 2018年

１．発表者名 Takashi Inui and Yuki Nakano
２．発表標題 An Analysis of Japanese Named Entity Recognizer Specialized for Person and Organization Entities
３．学会等名 The 22nd International Conference on Asian Language Processing（国際学会）
４．発表年 2018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