様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月１３日現在

研究種目：若手研究（Ｂ）
研究期間：2007～2009
課題番号：１９７００８９
研究課題名（和文）　文書データに出現する地名の意味認識システムの研究
研究課題名（英文）　A Study of Automatic Recognition Mechanisms of Meanings of Location Names in Document Databases

研究代表者
細川　宜秀　（HOSOKAWA　YOSHIHIDE）
群馬大学・大学院工学研究科・講師
研究者番号：５０３１２８３０

研究成果の概要（和文）：我々は，これまでに，文書データに出現する地名表現をそれが指すランドマークの緯度経度に翻訳するための技術（ジオ・コーディング技術）追求を行ってきた．その技術の特徴は，空間的文脈認識を伴って地名の指すランドマークの緯度経度を自動算出することにある．ここで，空間的文脈とは，説明文を構成する語群のうち，文書に含まれる地名表現が指し示す意味（緯度経度）を特定するのに貢献する語群を表す．しかしながら，我々のジオ・コーディング技術は，ランドマークに関する知識メタデータベース（ランドマーク・メタデータベース）を前提とするため，そのメタデータベースに登録されていないランドマークの緯度経度を指す地名を含む文書データを翻訳対象外としてきた．

本研究開発では，文書データベースを対象としたランドマーク・メタデータベースを自動生成するためのシステムの実現方式を実現した．本実現方式の主要な特徴は次の点にある：文書データベースから得られるランドマーク-単語間共起関係に基づいたランドマーク・メタデータ自動生成メカニズムの実現．これにより，先行研究で実現したジオ・コーディング技術の適用範囲を拡大することが可能になる．つまり，地理空間上に自動配置可能な文書数を大幅に増大させる．実験により，本実現方式の妥当性を明らかにした．

研究成果の概要（英文）：We developed a context-dependent geocoding system in our previous work. Geocoding systems are designed for translating location names to corresponding geocodes. The main feature of our geocoding system is to recognize the meanings of ambiguous location representations by handling the spatial contexts of the representations. We define a spatial context as a set of terms strongly relating to a single meaning of an ambiguous location representation. Our geocoding system was implemented to use a special metadatabase maintaining spatial contexts of landmarks.

In this study, we have developed a new implementation method for generating the metadatabase. The main feature of our method is to extract the spatial contexts of landmarks by handle the co-occurrence of the landmark names and feature words in a text database. Thus, our geocoding system can be applied to various text databases. We have demonstrated the feasibility and effectiveness of our method through several experiments.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	900,000	0	900,000
２００８年度	1,300,000	390,000	1,690,000
２００９年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,300,000	720,000	4,020,000

研究分野： 意味認識
科研費の分科・細目： 情報学・メディア情報学・データベース
キーワード： 意味認識，文書検索，位置情報システム，空間的文脈

１．研究開始当初の背景

文書データ中に含まれる地名の自動抽出，ならびに，該当する緯度経度の算出に関する研究とは，空間メディアと言語メディアを対象としたメディア交換技術に関わる学術分野に属する研究テーマである．そして，人の空間認知機能をサポートするための IT 技術確立という点において当該分野が IT 技術の発展に本質的であるにもかかわらず，空間メディアが画像メディアや音楽メディアに比べ目立たないこともあり，世界的に見ても他のメディア交換技術と比較して地味に発展してきた経緯を持つ．しかしながら，モバイル・コンピューティング技術の発展・普及，ならびに，Google Map APIs などのインターネットにおける地図と文書データの統合利用環境の出現により，検索者の現在位置，あるいは興味ある位置周辺に関連する情報収集のニーズが高まってきたこともあり，空間メディアと言語メディアの交換技術の価値が認識され始め，世界的にも活発化しつつある分野となってきた．

本研究課題が解決を目指す問題とは，文書データ中に含まれる地名の意味を認識する機能を実現することである．例えば，「王監督は，東京都内の病院に入院した．」という文書データに「東京」という地名が含まれているが，この「東京」が指す意味（この例では東京都新宿区にある慶応大病院）の緯度経度を算出することが本研究課題で取り組む問題である．この問題が解決されると，例文のような住所や緯度経度を含まないが，地名を含む文書データを実世界の該当する位置に自動的に配置することが可能になる．これにより，これまでは Web という実世界とは別の空間を検索対象としてきたが，それに加えて実世界上に蓄積された文書データの検索が可能になる．すなわち，IT 技術の利用場面が，従来の屋内活動（机の前に座して Web にアクセスする活動）に加え，検索者の屋外活動（移動を伴いながら実世界上に蓄積された文書データにアクセスする活動）に拡大される．特に，検索者の周辺に関する正確な情報獲得技術の確立は，検索者の次の社会活動のための意思決定を良質なものとするので，本研究課題の社会的インパクトは非常に大きいと言える．

当該分野初期に提案された手法は，緯度経度が格納された住所録を照合するものであるが，この場合，「東京」という文字列を含むすべての緯度経度（東京都に建設されたすべての建物の緯度経度はその一部である）が検索されてしまう．そこで，現在までに，地名の指す意味を特定することを目的とした次の２種類の先行研究が国内外で実施されてきた：

（先行研究１）「東京」が指す意味を特定するための語（(先行研究-１）ではこの語群を「文脈」と呼ぶ. 例文, 東京の後の「都」が文脈として扱われる.）からその意味を推測する方法

（先行研究２）意味が確定している他の地名から「東京」の意味を推測する方法

（先行研究１）では，「都」という文脈から東京都であることが推測されるが，東京都内のどの建物かまでは特定できない．すなわち，市などのある程度の広域レベルでの意味認識は実現されているが，詳細レベルでの意味認識は実現されていない.（先行研究２）は，２つ以上の地名が出現する文書データでないと適用困難である（すなわち，ここで挙げた例文は適用外である）という問題点を抱えている．

一方，自然言語処理分野において，文書データ中の地名を抽出するという研究が国内外でなされてきた．地名抽出に関して設定された解決すべき問題とは，地名表記の揺らぎ（主に略名）を吸収することであった．したがって，実現された主要な手法は，事前に与えた地名の表記フォーマットに従っているか否かを判定するもの（地名の構文との照合機能）であった．地名の意味認識機能の実現は達成されていないが，これらの成果は，空間メディア-言語メディア交換技術の第 1 ステップとして位置づけられる文書データからの地名抽出に使われている．

本研究課題代表者は，国内外のこれらの成果とは全くことなるアプローチにより「東京」の意味を特定する手法（ジオ・コーディング）を開発した．このアプローチは，スポーツファンが例文のような地名を含むスポーツ記事を読んだ際その地名の意味を特定できるが，スポーツに詳しくない人は地名の意味を特定できない様子を日常よく観察できることから着想したものである．すなわち，スポーツファンのみが持っている知識を使って地名の意味を特定しているという仮説に基づき実現した． その大まかな実行手順は，（慶応大病院，「王監督，入院」）のよう

な，スポーツファンの知識に対応する地名とその特徴の対応表を作成し，入力された文書データと地名の特徴との類似度を計量することによって意味を確定するものである．しかしながら，本技術をウェブなどの一般文書情報源に適用するための技術開発は未実施であった．

２．研究の目的

我々は，これまでに，文書データに出現する地名表現をそれが指すランドマークの緯度経度に翻訳するための技術（ジオ・コーディング技術）追求を行ってきた．その技術の特徴は，空間的文脈認識を伴って地名の指すランドマークの緯度経度を自動算出することにある．ここで，空間的文脈とは，説明文を構成する語群のうち，文書に含まれる地名表現が指し示す意味（緯度経度）を特定するのに貢献する語群を表す．しかしながら，我々のジオ・コーディング技術は，ランドマークに関する知識メタデータベース（ランドマーク・メタデータベース）を前提とするため，そのメタデータベースに登録されていないランドマークの緯度経度を指す地名を含む文書データを翻訳対象外としてきた．

本研究課題最大の目標は，そのランドマーク・メタデータベースの自動生成技術を実現することにある．この技術は，本研究代表者が先に開発したジオ・コーディング技術を一般文書に適用するのに必須となる技術の一つとして位置づけられる．

３．研究の方法

本研究課題が対象とする文書情報源として，新聞記事データベースを設定する．その理由は，次のとおりである．

（理由-１）本研究代表者が先に開発したジオ・コーディング技術を一般文書に適用するには，本研究課題として設定したランドマーク・メタデータベースの自動生成技術に加え，文書中の地名の意味を確定するのに貢献する空間的文脈の自動認識技術を確立する必要がある．しかしながら，空間的文脈の自動認識技術は，現在までに実現されておらず，また，一般に文書ごとに出現パターン（記述のされ方）が異なるため，本研究課題において，一般文書全般を我々のジオ・コーディング技術に適用することは困難である．一方，新聞記事は，時事速報を目的としているため，事象の空間的文脈が見出しや第１文に出現しやすい傾向にある．そこで，本研究課題では，空間的文脈の出現パターンが比較的固定化されている新聞記事データベースを対象にランドマーク・メタデータベースの自動生成技術の確立を目指す．

４．研究成果

次の特徴を有するランドマーク・メタデータベース自動生成方式を実現した．

（特徴-1）既存文書検索技術に基づいた非地理的特徴自動抽出機能の実現

（特徴-2）翻訳技術に適した非地理的特徴自動抽出機能の実現： 翻訳技術とは，正解が１つしかない入力キーワードに対して，その正解を第１位にランクづけできる検索技術ととらえることができる。したがって，複数ランドマークに関連する非地理的特徴を弱める効果をもたらす Inverse Landmark Frequency (ILF) を提案する。

$$ILF(t)=\log\left(\frac{|L|}{\left|\bigcup_{l_k\in L}\{(t,l_k)\mid d\in R\ (t\subseteq d\wedge l_k\subseteq d)\}\right|}\right)+1$$

ここで，L は，ランドマーク名集合を表す．l_k は，ランドマーク名を表す．R は，文書集合を表す．d は，ある 1 文書を構成する単語集合を表す．t は，単語を表す．

（特徴-3）複数新聞記事からのランドマーク非地理的特徴収集，合成機能の実現

次の手順によって，ランドマーク・メタデータベースを自動生成する．

（手順-1）文書からランドマーク名を抽出：これは，文書を構成する単語と地図データベースに含まれるランドマーク名との間の完全一致照合により行われる．

（手順-2）文書を名詞の列に変換：手順-1 の結果，その文書にランドマーク名が含まれていた場合，その文書を構成する単語群は，そのランドマークを特徴づける語（特徴語）の候補（特徴語候補）となりえる．また，本研究では，ランドマークの特徴語候補を名詞に限定する．

（手順-3）文書に含まれる各特徴語候補の特徴量を計算し，その値を要素とするランドマークの非地理的特徴語候補ベクトルを生成：本研究では，この特徴量計算を ILF に基づいて行う．

（手順-4）ベクトルの合成：あるランドマークを含む文書は複数存在する．そこで，各文書から抽出された特徴語候補ごとの特徴量を足し，そのランドマークに対する特徴語候補の特徴量を決定する．

（手順-5）特徴語候補群から，特徴量の大きな上位 m 候補をそのランドマークの特徴語として抽出

提案方式の妥当性を明らかにするために，その評価実験を行った．具体的には，手順-3 に既存文書検索において用いられているものを適用した方式（既存方式）によって自動生成されたランドマーク・メタデータの質に対する ILF を適用した提案方式によって自動

生成されたランドマーク・メタデータベースの質の良さを示す．なお，ランドマーク・メタデータベースの質は，我々が先行研究で実現した空間的文脈認識を伴うジオ・コーディング方式に適用し，その翻訳精度によって評価した．その結果は次のとおりである．

手順-3 に適用した方式	翻訳精度
ILF(提案)	0.73
IDF-ILF(提案)	0.76
TF-ILF(提案)	0.69
TF-IDF-ILF(提案)	0.69
BIN(既存)	0.73
IDF(既存)	0.75
TF(既存)	0.60
TF-IDF(既存)	0.69

この結果より，提案 ILF と既存 IDF を掛け合わせた ILF-IDF による手順-3 の実現方式の優位性を明らかにした．すなわち，本研究で実現したランドマーク・メタデータ自動生成方式の妥当性を明らかにした．なお，性能低下要因については，本稿において割愛させていただく．(それについては，学会発表①に記述されているので，そちらを参照されたい.)

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　0 件）

〔学会発表〕（計　3 件）

① 近藤好洋，細川宜秀，新聞記事データベースを対象とした空間的文脈認識を伴うジオ・コーディングのためのランドマーク・メタデータベースの自動生成機能の実現方式，電子情報通信学会 第 2 回データ工学と情報マネジメントに関するフォーラム(DEIM2010)，閲読有，2010.3.1，淡路夢舞台国際会議場（兵庫県淡路市）
② 細川宜秀，図への文書自動配置機能の地域内情報発信システムへの適性評価，電子情報通信学会 第 2 回データ工学と情報マネジメントに関するフォーラム(DEIM2010)，閲読有，2010.3.1，淡路夢舞台国際会議場（兵庫県淡路市）
③ 中澤優一郎，細川宜秀，永島和矩：空間的関係の近似化を伴う周辺情報提示機構の実現方式，電子情報通信学会 第 2 回データ工学と情報マネジメントに関するフォーラム(DEIM2010)，閲読有，2010.2.28 舞台国際会議場（兵庫県淡路市）

〔その他〕
ホームページ等
http://www.dbsp.cs.gunma-u.ac.jp

６．研究組織
(1)研究代表者
細川　宜秀（HOSOKAWA　YOSHIHIDE）
群馬大学・大学院工学研究科・講師
研究者番号：５０３１２８３０

(2)研究分担者
（　　）
研究者番号：

(3)連携研究者
（　　）
研究者番号：