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

近藤　好洋† 　細川　宜秀††

† 群馬大学工学部情報工学科
†† 群馬大学大学院工学研究科

あらまし　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．本稿では，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提案する．提案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，提案方式の妥当性を明らかにする．
キーワード　ジオ・コーディング，メタデータ自動生成，地理情報システム

An Implementation Method of An Automatic Landmark Metadata Extraction System for Context-Dependent Geocoding for A Text-based News Database

Yoshihiro KONDO† and Yoshihide HOSOKAWA††

† Department of Computer Science, Gunma University
†† Graduate School of Engineering, Gunma University

Abstract　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 paper, we present 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 clarify the feasibility and effextiveness of our method through several experiments.
Key words　Geocoding, Automatic Metadata Extraction, Geographical Information Systems

1. はじめに

現在，次の 4 項目を根拠に，コンピュータ・ネットワーク上にある文書情報源からの，緯度経度が関連付けられた情報の獲得ニーズが増大している．

（1） 無線ネットワーク技術，位置センシング技術，端末小型化技術の発展・普及は，コンピュータ・ネットワーク上に蓄積された地域に関する情報（地域情報）を獲得するためのハー

ドウェア基盤構築に貢献してきた．

（2） 空間データベース分野において，緯度経度表現を対象とした情報検索・統合のためのソフトウェア技術が開発され，地理情報システムを代表的な応用として実用されてきた．

（3） Google Maps API などのインタネット地図サービスは，コンピュータ・ネットワーク上で緯度経度付きの情報の量を飛躍的に増大させた．つまり，このサービスは，多数のコンピュータ・ネットワーク利用者に緯度経度付き情報の獲得の有用性を気付かせるトリガとなった．

（4） 広域コンピュータ・ネットワーク技術とウェブ技術の発展・普及により，コンピュータ・ネットワーク上に地球規模の文書情報源が構築された．そのボリュームは，現在も増大している．

しかしながら，次の課題がそのニーズ実現を妨げている：**(課題) コンピュータ・ネットワーク上の文書情報源は，地名などの空間テキスト表現を含むが，その地名が指す緯度経度が関連付けられていない文書データを多数含んでいる．**

我々は，これまでに，この課題に対し，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．この技術の特徴は，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）を前提に動作するため，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．

実現方式の特徴は次の 2 点にある．

特徴-1： 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に関する空間的文脈自動抽出メカニズムの実現

（地名を対象とした）翻訳システム実現要件は，次の 2 項目を同時に満たすことにある：(要件-1) システムが返す翻訳結果が少数であること，ならびに，(要件-2) その結果に必ず正解を含んでいること．ここで，文書データに出現する地名の多くが，実空間上の唯一のランドマークを指す（正解となりえるランドマーク数は 1 である）ことを考慮すると，その 2 要件に示した少数とは，多くの場合 1 となる．つまり，本稿が対象とする課題とは，正解が 1 つしかないクエリに対して，その正解を常に最上位結果として返す検索技術を開発することにある．

そこで，我々は，どのような空間的文脈が与えられた場合においても（要件-1）を満たす 1 実現方法として，ランドマークに関連する空間的文脈を，互いに可能な限り共通の特徴を持たないように自動抽出するためのメカニズム実現を目指す．具体的には，文書データベースにおいて，複数のランドマークと共起する単語群をランドマークの空間的文脈として採用しないようにするためのメカニズム実現を目指す．

特徴-2： 新聞記事データベースを対象としたランドマークに関する空間的文脈自動抽出メカニズムの実現

多くの新聞記事は，あるランドマークで発生した特定の時事イベントやオブジェクトに関する記述に集中している． つまり，単一の新聞記事は，そのランドマークの 1 側面を記述している文書とみなせる．したがって，あるランドマークの空間的文脈を漏れなく抽出するには，そのランドマーク名を含むすべての新聞記事からその空間的文脈の部分を抽出し，それらの部分を合成するためのメカニズムを実現することが本質的である．これより，翻訳時にどのような空間的文脈が与えられても，それに適合するランドマークを漏れなく返すジオ・コーディングが実現される．すなわち，特徴-1 において示した（要件-2）を満足するジオ・コーディングが実現される．

実験により，提案方式の妥当性を明らかにする．

なお，我々のジオ・コーディング技術の具体的な活用シーンとして，地図を介した文書検索に加え，緯度経度を介した構造化データベース（GIS はその具体的な実用形態）と文書データベースの連結による情報獲得，さらには，文献 [6] に示される地域情報源構築がある．

2. 関 連 研 究

文書データに含まれる概念の特徴，ならびに，概念間の関連性を自動算出する手法が多数研究されている．文献 [8] はクラスタリング技術を用いてウェブページ集合をクラスタに分割し，各クラスタから TF・IDF を用いて地理オブジェクトに関する特徴キーワード抽出を行うものである．文献 [3] は，Web ページの内容とページレイアウトから用語説明を抽出し，分野ごとに分類する方法を示している．文献 [3] は，文書表現を用いて用語説明を抽出し，確率モデルを用いて用語説明を分類する．文献 [11] は，Web ページの内容から用語説明文を抽出し，入力用語に適した説明文を提示する方法を示している．文献 [11] は，文書表現を用いて用語説明を抽出し，ベクトル空間モデルとコサイン類似度を用いて入力用語に適した説明文を見つける．文献 [7] は，Wikipedia におけるページ間リンク共起関係から，関連する概念を算出するものである．

従来のジオ・コーディングは地理的特徴のみを用いて，地名表現を緯度経度に翻訳する [1] [9] [10] [12].

3. 提 案 方 式

本節では，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・メタデータベース自動生成機能の実現方式 (提案方式) について述べる．提案方式によって，先行研究 [4] [5] 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．

3.1 データ構造

提案方式は，次の 3 データベース上に定義される．

- ランドマーク・メタデータベース
- 新聞記事データベース
- 地図データベース

ランドマーク・メタデータベースは，先行研究で実現した空間的文脈認識を伴うジオ・コーディングを実現するための本質

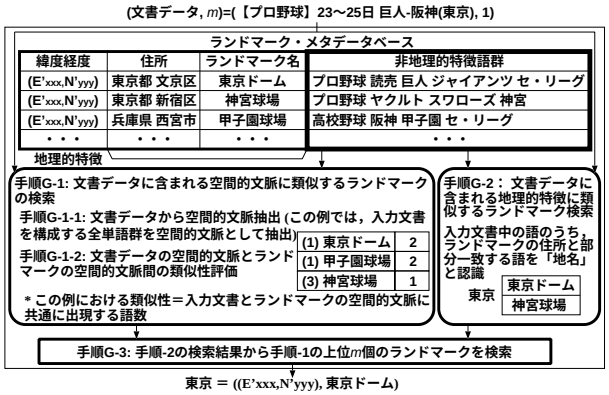

図 1　空間的文脈認識を伴うジオ・コーディングの実行例 ($m = 1$; m は出力個数を表す．)

的なデータベースとして位置付けられる．このデータベースは，ランドマークの名前，その住所，その緯度経度，ならびに，その非地理的特徴語群からなる 4 つ組集合として定義される．ここで，地理的特徴とは，空間的文脈の内単語にて緯度経度を特定できる語と定義する．非地理的特徴とは，空間的文脈の内，地理的特徴以外の語と定義する．

新聞記事データベースと地図データベースは，ランドマーク・メタデータベースの自動生成のための素材データ群として使用される．新聞記事データベースは，新聞記事を表すプレーン・テキストの集合として定義される．地図データベースは，ランドマーク名，その住所，ならびに，その緯度経度からなる 3 つ組集合として定義される．

3.2　空間的文脈認識を伴うジオ・コーディング実現手法

本節では，先行研究 [4] において実現したジオ・コーディングの実現方式を示す．図 1 はその実行例を表す．この例は，文書データに含まれる地名 (東京) を地名が指す緯度経度 (東京ドーム) への翻訳を行うものである．

手順は次のとおりである．

手順 G-1：　文書データに含まれる空間的文脈に類似するランドマーク検索を行う．図 1 の手順 G-1 では，' 東京ドーム'，' 神宮球場'，' 甲子園球場' の順にランドマークが検索される．文書データと相関が大きいランドマーク順にランドマークが検索される．

手順 G-1-1：　文書データから空間的文脈抽出する．具体的には，文書データを形態素に分割し，その形態素語群から空間的文脈を形成する．なお，形態素解析には，主要な形態素解析器である ChaSen [2] を用いた．

図 1 は，入力文書を構成する全語群を空間的文脈として取り扱う方法を示している．現在の本プロトタイプ・システムは，手順 G-2 において抽出された地名毎に，(i) 新聞記事の見出し，(ii) 第 1 文，ならびに，(iii) その地名の前後 3 名詞を空間的文脈として抽出するように実装されている．これは，新聞記事において，見出しと第 1 文に事象の発生時刻と発生場所に関する記述がよく出現することに基づく．

手順 G-1-2：　文書データに含まれる空間的文脈とランドマークの空間的文脈間の類似性評価を行う．なお，ランドマークの空間的文脈を，ランドマーク名，住所，ならびに，非地理的特徴語群を連結することによって導出する．

図 1 は，それらの間の類似性を文書データとランドマークの空間的文脈に共通に出現する語数により算出する様子を示している．現在の本プロトタイプ・システムは，文書データとランドマークの空間的文脈の類似性評価に，主要な文書検索モデルとして知られるベクトル空間モデルを採用している．各空間的文脈は，形態素を要素とする特徴ベクトルとして表現される．特徴ベクトルの各要素値は，対応する形態素の TF·IDF 値により表現される．空間的文脈間の類似性は，それらの特徴ベクトルの内積値として算出される．

手順 G-2：　文書データに含まれる地理的特徴に類似するランドマーク検索を検索する．

図 1 に示すように，本システムは，' 東京' を含む住所文字列を有するランドマークをランドマーク・メタデータベースから検索する．その結果として，' 東京ドーム' と' 神宮球場' が翻訳結果候補として検索される．

手順 G-3：　手順 G-2 の検索結果から手順 G-1 の上位 m 個のランドマークを検索する．図 1 は，本システムが入力文暑中の '東京' を '東京ドーム' に翻訳する様子を表す．

3.3　新聞記事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方式

3.3.1　空間的文脈を構成する語の選定方式

我々は，地名の緯度経度への翻訳技術を実現するにあたり，特定のランドマークと結びつきの強い語を積極的に取り扱うことにより，与えられた空間的文脈に即した翻訳結果候補を絞り込むことが可能であると考える．ランドマーク名と単語間の共起回数は，ランドマークと単語間の結びつきの強さを表現する 1 指標とみなせる．すなわち，単一，あるいは，ごく少数のランドマークと共起する単語は，それらのランドマークと強い結びつきがあるとみなせる．

そこで，我々は，文書データベースにおけるランドマーク名-単語間の共起関係に基づいた翻訳結果候補絞り込み指標 ILF(Inverse Landmark Frequency) を提案する．

$$ILF(t) = log(\frac{|L|}{lf^{Dt}(t)}) + 1,$$

$$lf^{Dt} = \left| \bigcup_{l_k \in L} \{(t, l_k) | \exists d \in R(t \subseteqq d \wedge l_k \subseteqq d)\} \right|$$

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．

ILF は，文書検索システムにおいてよく用いられる IDF (Inverse Document Frequency) の考え方に基づいたものである．ただし，(t, l_k) が繰り返し出現する特定の文書の影響度が反映されないよう，単一文書から得られるランドマーク名-単語間の共起関係はたかだか 1 とする．なお，ILF は，図 2 に示す lf^{Dt} の 1 活用形態として位置付けられる．

提案方式として，ILF と文書検索システムにおける有効な

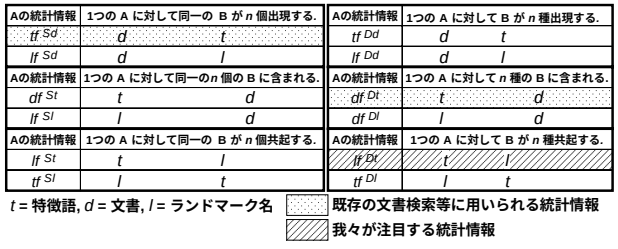

図 2　ランドマーク-単語-文書間の統計情報の種類

図 3　提案方式によるランドマーク・メタデータ自動生成手順

指標として位置付けられる TF (Term Frequency) と IDF を組み合わせた次の 4 重み付け方式に基づいたメタデータ自動生成方式を実現する.

- ILF
- $IDF{\cdot}ILF$
- $TF{\cdot}ILF$
- $TF{\cdot}IDF{\cdot}ILF$

なお，TF は図 2 に示す tf^{Sd} の，IDF は df^{Dt} の活用形態に分類される.

3.3.2　新聞記事データベースを対象としたランドマーク・メタデータ自動生成方式

本節において，提案するランドマーク・メタデータベース自動生成システムの実行手順を示す．図 3 はその実行例を表す.

手順 M-1： 文書からランドマーク名を抽出する．文書を構成する名詞のうち，地図データベースのランドマーク名と完全一致するものをランドマーク名と判定する．文書からの名詞抽出手順は，次のとおりである：(M-1-1) 文書を形態素に分割する．(M-1-2) 名詞に分類される形態素の並びを連結する．現在のプロトタイプシステムの形態素解析は，ChaSen を用いて行う.

手順 M-2： 文書からランドマーク名と空間的文脈候補の対を抽出する．ここでは，空間的文脈候補の語として，形態素の品詞が「名詞」と「未知語」のものを抽出する.

手順 M-3： 空間的文脈候補の各語の重みを計算し，その値を要素とする空間的文脈候補ベクトルを生成する．各要素値は，本稿第 3.3.1 節において示した重み付け方式によって算出される.

手順 M-4： ランドマークの空間的文脈候補ベクトルを合成する．ここで，ランドマーク l が文書データベース R 中の k 個の文書に出現するとする．さらに，k 個の文書から n 個の名詞と未知語が抽出されたとする．このとき，ランドマーク l に関する合成された空間的文脈候補ベクトルは，次の式により表現される.

$$(\sum_{i=1}^{k} t_{(i,1)}, \cdots, \sum_{i=1}^{k} t_{(i,j)}, \cdots, \sum_{i=1}^{k} t_{(i,n)})$$

手順 M-5： 合成された空間的文脈候補ベクトルから，要素値の大きい上位 m 個の語を選択する．現在のプロトタイプにおいて m に 10 を与えた.

手順 M-6： 抽出された空間的文脈候補をランドマーク・メタデータベースの非地理的特徴語群属性値として登録する．なお，提案方式では，自動抽出された非地理的特徴語群から，地理的特徴語に一致する語を非地理的特徴語として扱うため，非地理的特徴語群からのその語の除去は行わない．例えば，図 3 の「東京ドーム」を非地理的特徴語群から除去しない．ちなみに，「東京ドーム」は，ランドマーク名とそのランドマークを管理する会社名に一致する．前者の場合，地理的特徴語になるが，後者の場合，非地理的特徴語になる．この場合には，ランドマークとしての東京ドーム内に会社「東京ドーム」が所在するため問題は生じないが，東京ドームを管理する会社が東京ドームではなく，都心のビル内に所在しても何ら不思議ないと言える．なお，提案方式において，その区別を行わない理由は，難課題であること，ならびに，本稿の範囲を越えることによる.

4. 実　　験

提案方式の妥当性を明らかにするために，次の 4 実験を行った．それらの目的と内容は，次のとおりである.

実験-1　本実験の目的は，提案方式によって自動生成されたランドマーク・メタデータの質を評価することにある．本実験では，提案方式と人手により生成されたランドマーク・メタデータベースを用いたジオ・コーディング翻訳精度を比較することによって，その評価を実施する.

実験-2　本実験の目的は，本稿第 3.3 節で述べた ILF に基づいたランドマーク・メタデータ自動生成方式の妥当性を評価することにある．本実験において，提案方式と，ILF を使用しないランドマーク・メタデータ自動生成方式によって導出されたランドマーク・メタデータの質を比較を行う．ここで，ILF を使用しないランドマーク・メタデータ自動生成方式とは，文書検索システムにおいてよく用いられている単語の特徴量に基づいたランドマーク・メタデータ自動生成方式である．具体的には，

表 1　各実験における比較対象方式 (添え字 $_m$ と $_s$ は，それぞれ，複数，ならびに，単一新聞記事から，単一ランドマークに関するランドマーク・メタデータを自動構成する方式を表す．)

	提案方式	比較対象方式
実験-1	ILF_m, $TF \cdot IDF \cdot ILF_m$ $IDF \cdot ILF_m$, $TF \cdot ILF_m$	M(人手による空間的文脈抽出方式)
実験-2	ILF_m, $TF \cdot IDF \cdot ILF_m$ $IDF \cdot ILF_m$, $TF \cdot ILF_m$	BIN_m, IDF_m TF_m, $TF \cdot IDF_m$
実験-3	ILF_m, $IDF \cdot ILF_m$ $TF \cdot ILF_m$ $TF \cdot IDF \cdot ILF_m$	BIN_s, TF_s ,$TF \cdot IDF_s$, IDF_s, ILF_s, $IDF \cdot ILF_s$ $TF \cdot ILF_s$ $TF \cdot IDF \cdot ILF_s$
実験-4	ILF_m, $TF \cdot IDF \cdot ILF_m$ $IDF \cdot ILF_m$, $TF \cdot ILF_m$	M

1 文書における単語出現の有無 (二進重み；BINary weighting)，単語の出現回数 (Term Frequency)，ならびに，ある 1 単語が出現する文書数 (Document Frequency) を組み合わせ定義される 4 方式 (BIN， TF， IDF， $TF \cdot IDF$) を比較対象方式として採用する．

ランドマーク・メタデータの質に関する評価を，提案方式と比較対象方式によって作成されたランドマーク・メタデータベースを我々のジオ・コーディング・システムに適用し，その翻訳精度を比較することによって行う．

実験-3　本実験の目的は，複数の新聞記事からランドマーク・メタデータベースを自動構成すする提案方式の妥当性を明らかにすることにある．本実験では，提案方式と，単一の新聞記事からのランドマーク・メタデータ自動生成方式により生成されたランドマーク・メタデータベースを用いたジオ・コーディング翻訳精度を比較することにより，提案方式によって生成されたランドマーク・メタデータの質を評価する．

実験-4　本実験の目的は，提案方式の大規模文書情報源への適性を明らかにすることにある．この適性を，地図データベースに含まれるランドマーク数に対する，新聞記事データベースからランドマーク・メタデータを自動構成されたランドマーク数の割合（カバー率）によって評価する．

表 1 は，各実験における比較対象方式をまとめたものである．本稿では，実験結果における各方式の性能をこの表の記号を用いて提示する．

4.1 実験環境

4.1.1 人手によるランドマーク・メタデータベース

本実験おいて,「MapInfo スタンダード道路地図 2003 施設ポイント・データ」に収録されている 568,877 施設を空間的文脈抽出対象ランドマークとした．また，このデータは，提案方式が使用する地図データベースとして活用される．

人手による，新聞記事からの，ランドマーク・メタデータ作成作業を次の手順により実施した．

作業-1：　約 10 名の作業者を雇用し，毎日新聞 CD-ROM データ集 (2002 度版) の主にスポーツに関する記事から，ランドマーク・メタデータを抽出した．この作業を半年間 (2003 年 8 月 ~2004 年 1 月) 実施した．

作業-2：　我々は，作業者によって抽出されたランドマーク・メタデータの表記ミスを修正した．

作業-3：　我々は，我々の眼から見て類似している空間的文脈を持つ同一ランドマークのメタデータを 1 つに合成した．ただし，異なる空間的文脈が割り当てられた同一ランドマークをまとめることはしなかった．これは，多義性を持つランドマークが実世界に存在することによる．

これより，我々は，4,801 のランドマークに関する 4,881 行からなるランドマーク・データベースを構築した．

4.1.2 正解集

ジオ・コーディングの翻訳精度を評価するために地名に関する正解集を作成した．この正解集は，毎日新聞 CD-ROM データ集 (2002 度版) のスポーツに関する 4,445 記事に含まれる地名，ランドマーク名，チーム名，組織名に対し，それが指すランドマーク（の緯度経度）を人手により割り当てたものである．正解集を構成する 4,445 記事の選定手順は次のとおりである：(選定手順-1) 毎日新聞 CD-ROM データ集 (2002 度版) のスポーツに関する記事（以後,「スポーツ記事」と略す.）からランダムに 4,052 記事を選択，(選定手順-2) 選択された記事数の少ないスポーツ分野の記事を人手により追加.

本実験をスポーツ記事に限定して実施する理由は次のとおりである：(理由-1) スポーツ分野は，野球やバレーボールなど多数の独立細分野に分かれており，異種分野を含んだランドマーク・メタデータベース構築の弊害を検証するには十分である．(理由-2) プライバシー問題に関わる地名に関し，著者らが正解ランドマークを割り当てることができなかった．具体的には，密会や個人宅を指すランドマークを特定することが困難であった．この理由に該当する記事として，政治と生活に関する記事が挙げられる．(理由-3) 市販の地図データベースには，主として公共ランドマークのみが収録されている．言い換えれば，市販の地図データベースには，政治家事務所や生活面に登場する個別の自宅，さらには，経済面に出現する中小企業がランドマークとして登録されていない．自動車事故など道路上の地点も事故現場として地図データベースに登録されていない．結果として，それらを指す地名の翻訳を実施できない．

4.2 実験方法

本実験では，次の 2 評価視点に基づいたランドマーク・メタデータの質評価を行う：(評価視点-A) 人手作成によるランドマーク・メタデータとの比較によるランドマーク・メタデータ質評価，ならびに，(評価視点-B) 異なるランドマーク・メタデータ間の干渉によるジオ・コーディング性能低下を抑制する良質ランドマーク・メタデータ生成．各視点に応じて，メタデータ自動生成方式と人手によって作成されるランドマーク・メタデータベースを再構成する．

メタデータ・セット-A：　本セットは，評価視点-A から提案方式を評価するのに使用される．具体的には，次の 2 条件を同時に満たすランドマークに関するメタデータのみを含むように再構成したランドマーク・メタデータベースである：(条件-1) 人手，ならびに，自動生成方式によって生成されたランドマーク・メタデータベースに共通に出現するランドマーク，ならび

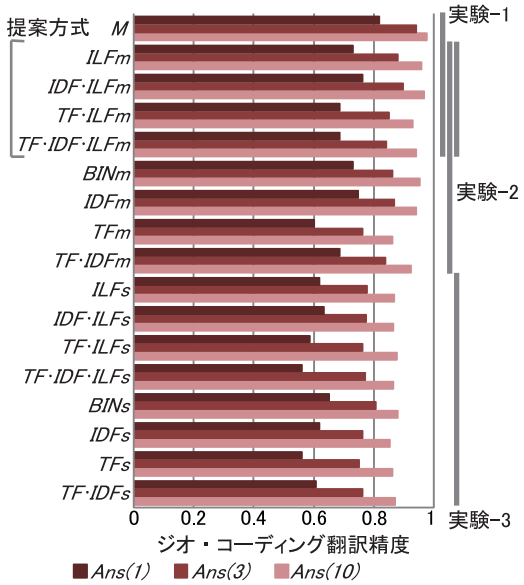

図 4　メタデータ・セット-A に対する実験-1 ～ 実験-3 の結果

に，(条件-2) 正解集を構成するランドマーク．

結果として，217 の共通ランドマークに関するランドマーク・メタデータベースを構築した．メタデータ・セット-A に対し，正解集に含まれる 1,684 地名 (それを含む新聞記事数 = 934) のジオ・コーディングを実施した．

メタデータ・セット-B： 本セットは，評価視点-B から提案方式を評価するのに使用される．これは，作成された生来のランドマーク・メタデータベースである．すなわち，人手によって作成されたランドマーク・データベースは，4,801 のランドマークに関するメタデータからなる．メタデータ自動生成方式によって生成されたランドマーク・データベースは，88,702 のランドマークに関するメタデータからなる．

ジオ・コーディング翻訳精度 $Ans(i)$ を次の式により算出する．

$$Ans(i) = \frac{A}{L}$$

ここで，A は，ジオ・コーディング結果の上位 i 位までに正解ランドマークが出現する地名数を表す．L は，実験に使用した総地名数を表す．本実験において，i を 1, 3, 10 と変化させた．

4.3　実験結果と考察

4.3.1　実験-1 ～ 実験-3 の結果と考察: 空間的文脈の質評価

図 4 は，メタデータ・セット-A に対するジオ・コーディングの翻訳精度（実験-1 ～ 実験-3 の結果）を表す．図 5 は，記事ジャンル毎のメタデータ・セット-A に対するジオ・コーディングの翻訳精度 $Ans(3)$ を表す．ここで，縦軸は記事ジャンルを表す．括弧内の数字は，ジオ・コーディングを行った地名数を表す．表 2 は，各方式によって生成された 5 ランドマークの非地理的特徴を表す．

まず，図 4 より，実験-3 の結果として，複数の新聞記事から生成されたランドマーク・メタデータは，単一の新聞記事から生成されたそれと比べ良質であった．この理由は，各新聞記事から抽出されるランドマークの非地理的特徴がお互いに重複しない部分を持つことによる．その結果として，ある新聞記事

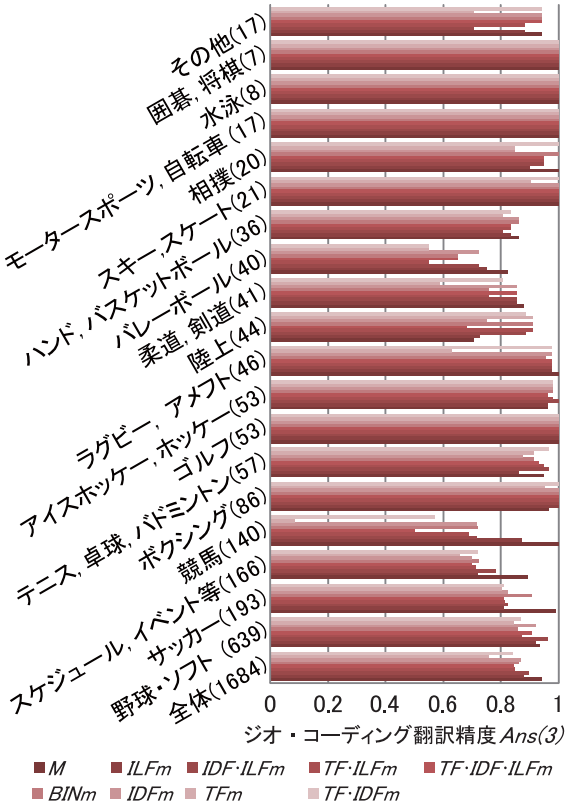

図 5　記事ジャンル毎のメタデータ・セット-A を使用したジオ・コーディング翻訳精度 $Ans(3)$

に含まれる地名がランドマーク-A を指す場合において，次の 2 語群間の共通項の少なさは，ジオ・コーディング翻訳精度を低下させる：(語群-1) その記事に含まれ，かつ，その地名がランドマーク-A を指すことを決定付ける語群，ならびに，(語群-2) 他の新聞記事から抽出されたランドマーク-A の非地理的特徴．これより，複数新聞記事からのランドマーク・メタデータ自動生成方式は，空間的文脈認識を伴うジオ・コーディングに適したランドマーク・メタデータ生成を行えることを明かにした．これ以後，人手によるランドマーク・メタデータ生成方式と，複数新聞記事からランドマーク・メタデータを自動生成する方式の比較に集中する．

図 4，図 5，ならびに，表 2 は，次の結果を示している．

（1） ILF_m と IDF_m を活用したメタデータ自動生成方式によるランドマーク・メタデータは，他方式のそれと比較して，良質であった．特に，メタデータ自動生成方式の中で，総合的には，$IDF{\cdot}ILF_m$ が最も良質のランドマーク・メタデータを抽出した (図 4)．この理由は，表 2 に示す「トヨタ自動車」の非地理的特徴に示すように，それらを組み合わせた活用は，単体での活用に比べ，数字のような多くの非地理的特徴の共通項除去に貢献することによる．ただし，特定の記事にのみ出現する人名やライバル・チーム名を含むようになる．この影響は，記事ジャンル毎のジオ・コーディング翻訳精度のばらつきに見られる．

また，M との比較において，$IDF{\cdot}ILF_m$ の $Ans(1)$ と $Ans(3)$ は，わずか 1 % の性能差であった．

（2） ジオ・コーディングが適用された回数が 100 を越えるジャンルのうち,「競馬」と「スケジュール，イベント等」を対象としたジオ・コーディング翻訳精度は，それ以外のジャンルを対象としたそれに比べ，低い値を示した (図 5)．この理由は，翻訳対象地名の空間的文脈としてそれを含む文書の「見出し」,「第 1 文」，ならびに，その地名前後 3 名詞と決め打ちしたことにある.「競馬」と「スケジュール・イベント等」の記事の特徴は，1 記事中に複数のスポーツ・イベントを含むことにある．しかしながら，それらの記事の「見出し」は，その記事において最も主張したいスポーツ・イベントに関連することが多い．さらに,「第 1 文」は，その記事において最初に記述されたスポーツ・イベントに関連することが多い．したがって，本実験のために設定した地名の空間的文脈と異なる空間的文脈を持つ地名に対するジオ・コーディングの翻訳精度は低下した．

（3） ジオ・コーディングが適用された回数が 50 未満のジャンルのうち,「陸上」,「柔道，剣道」,「バレーボール」,「ハンド，バスケットボール」を対象としたランドマークに関する非地理的特徴の質は良くなかった．(図 5) この理由は，複数ジャンルのスポーツ・イベントが開催されるランドマークに関する非地理的特徴に，その一部のジャンルに関連する語が含まれなかったことにある．例えば，表 2 に示す「中央体育館」(大阪市) に関して，ハンドボールやソフトテニスなどの試合が開催されたが，メタデータ自動生成方式が生成した中央体育館の非地理的特徴にその内容に関する語群が含まれなかった．

「陸上」を対象としたランドマーク・メタデータの質に関しても同様に言える．表 2 に示す「国立京都国際会館」は，全国高校駅伝の折り返し地点に位置するが，その非地理的特徴には，それに関するものが含まれなかった．

（4） TF 値を活用しないメタデータ自動生成方式は，TF 値を活用したそれに比べ，良質なランドマーク・メタデータ生成を達成した．この理由は，表 2 に示すように，TF 値を活用したメタデータ自動生成方式によって抽出された非地理的特徴に数字が多く含まれることにある．これは，スコアやタイムなど多くの数字を含むスポーツ記事を対象としたジオ・コーディング翻訳精度を低下させる主要因となった．

以上より，ILF_m の活用は，空間的文脈認識を伴う地名翻訳に効果があることが明かとなった．

4.3.2 実験-4 の結果と考察: 空間的文脈自動抽出方式の有用性

提案方式は，88,702 ランドマークに関するランドマーク・メタデータを自動生成した．この数は，人手によって作成したランドマーク数の 18.5 倍 ($=\frac{88,702}{4,801}$) に等しい．一方，この数は，地図データ上に定義されているランドマーク数の 0.16 倍 ($=\frac{88,702}{568,877}$) に過ぎない．これは，地図データに含まれるランドマークの名前を含む新聞記事が少ないことによる．この課題解決には，ランドマーク名に関する地図と新聞記事間のインターオペラビリティを達成する手法の実現が必須となる．なお，この手法は，本稿の範囲を越える．

以上より，新聞記事からのランドマーク・メタデータ自動生成に関するある程度の有用性を明かにした．

4.3.3 データベース・セット-B を用いた実験結果と考察

研究発表において，その結果と考察を示した．紙面の都合上，省略する．

5. ま と め

本稿では，文書データベースからランドマーク・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提案した．提案方式の特徴は次の点にまとめられる．(特徴-1) 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，ならびに，(特徴-2) 新聞記事データベースを対象としたランドマーク・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．提案方式の妥当性を示すための実験を行った．

今後の課題として，図 2 に示した重み付け手法のうち，本稿に示した方式において使用していないものを活用した新しいランドマーク・メタデータ生成方式の実現が挙げられる．

謝　　辞

本研究の一部は，科学研究費補助金若手研究 (B)(#19700089) によるものである．

文　　献

[1] Bakshi, R., Knoblock, C. A., Thakkar, S., Exploiting online sources to accurately geocode address, *Proc. 12th ACM International Workshop on Geographic Information Systems (ACM-GIS 2004)*, pp.194–203 (2004)

[2] ChaSen, `http://ChaSen.naist.jp/hiki/ChaSen`

[3] 藤井 敦, 石川 徹也, World Wide Web を用いた辞典知識情報の抽出と組織化, 電気情報通信学会論文誌 D-II Vol.J85-D-II No.2 pp.300-307 (2002 年 2 月)

[4] Hosokawa, Y., Takahashi,N., A context-dependent geocoding method for docment databases, Information Modelling and Knowledge Bases(IOS Press), Vol.16, pp.225–239(2005)

[5] 細川 宜秀, 高橋 直久, ドキュメント・データを対象としたジオ・コーディング手法, 情報処理学会研究報告. データベース・システム研究会報告 , p87-93(2003)

[6] 細川 宜秀, 地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) 論文集 (2010)

[7] 伊藤 雅弘, 中山浩太郎, 原 隆浩, 西尾章治朗, センテンスを考慮したリンク共起性解析による Wikipedia からの連想シソーラス辞書構築に関する一考察, DEWS2008, A3-1

[8] 中戸 隆一郎, 岩井原 瑞穂, ウェブ地域情報の自動要約のための特徴キーワード抽出, DEWS2005 , 5C-o3

[9] Silva, M., Martins, B., Chaves, M., Afonso, A., and Cardoso, N., Adding geographic scopes to web resources, *Computers, Environment and Urban Systems*, Vol.30, No.4, pp.378–399 (2006)

[10] Tezuka, T., Yokota, Y., Iwaihara, M., and Tanaka, K., Extraction of Cognitively-Significant Place Names and Regions from Web-Based Physical Proximity Co-occurrences, *Proc. 5th International Conference on Web Information Systems Engineering*, pp.113–124 (2004)

[11] 土田 正明, 松井 藤五郎, 大和田 勇人, WorldWideWeb を用いた辞典システムの構築, 人工知能学会全国大会論文集 18th 1A3-04 (2004)

[12] Wang, C., Xie, X., Wang, L., Lu, T., and Ma, W. Y., Detecting Geographic Locations from Web Resources, *Proc. the 2005 workshop on Geographic information retrieval (GIR'05)*, pp.17–24 (2005)

表 2　各方式によって作成された一部のランドマークの非地理的特徴

ランドマーク名	方式	非地理的特徴
京都競馬場	M	競馬 中央競馬 日曜競馬 日本中央競馬会 京都競馬 土曜競馬 芝 ダート
	ILF_m	京都競馬場 ＪＲＡ 日本中央競馬会 善雄 オッズ トランス 茶谷 ウインズ ファー 競馬
	$IDF{\cdot}ILF_m$	京都競馬場 善雄 オッズ ＪＲＡ 茶谷 日本中央競馬会 ウインズ トランス 千代崎 馬番
	$TF{\cdot}ILF_m$	ＪＲＡ １ 容疑 茶谷 オッズ カード ０ 京都競馬場 者 ２
	$TF{\cdot}IDF{\cdot}ILF_m$	ＪＲＡ 茶谷 オッズ 京都競 馬場 カード ウインズ 容疑 プリンター 競馬 善雄
	BIN_m	京都競馬場 １ ２ ９ ７ ０ ５ ６ 円 万
	IDF_m	京都競馬場 日本中央競馬会 ＪＲＡ 善雄 茶谷 トランス オッズ ファー 和男 千代崎
	TF_m	１ ０ ２ ３ 者 容疑 ５ ７ ９ ６
	$TF{\cdot}IDF_m$	ＪＲＡ 茶谷 カード 容疑 オッズ １ 京都競馬場 ０ 鈴木 者
甲子園球場	M	高校野球 日本高校野球連盟 第７４回センバツ高校野球 プロ野球 甲子園 阪神 アメリカンフットボール クラッシュボウル 第５７回毎日甲子園ボウル 全国高校野球選手権大会 関西学生野球 夏の甲子園 神 甲 第７４回選抜高校野球大会 タイガース プロ野球オープン戦 セ・リーグ 東西大学王座決定戦 タイガース
	ILF_m	球場 甲子園 野球 センバツ 阪神 試合 高校 阪神タイガース 監督 選手
	$IDF{\cdot}ILF_m$	球場 甲子園 センバツ 阪神タイガース 阪神 野球 試合 球児 タイガース 高校
	$TF{\cdot}ILF_m$	甲子園 阪神 ０ １ 野球 球場 ２ 監督 星野 センバツ
	$TF{\cdot}IDF{\cdot}ILF_m$	甲子園 阪神 球場 星野 センバツ 野球 監督 感動 選手 阪神タイガース
	BIN_m	球場 甲子園 ２ １ 日 ５ 野 球 ４ ０ ３
	IDF_m	球場 甲子園 野球 センバツ 阪神 高校 試合 阪神タイガース 監督 大会
	TF_m	０ １ ２ ３ ５ ４ 野球 甲子園 日 阪神
	$TF{\cdot}IDF_m$	甲子園 阪神 野球 球場 ０ 星野 監督 １ センバツ 選手
中央体育館 (大阪市)	M	高松宮杯男子第４５回・女子第３８回全日本学生ハンドボール選手権大会 バレーボール Ｖリーグ 第４７回全日本インドア・ソフトテニス選手権大会 ＷＢＣフライ級王座戦 プロボクシング 女子決勝R ワールドリーグ
	ILF_m	体育館 中央 ボクシング 大阪 グリーンツダ バレーボール 戦 市 クラティンデンジム 日
	$IDF{\cdot}ILF_m$	体育館 グリーンツダ クラティンデンジム エディタウンゼント ボクシング バレーボール ＷＢＣ レック タイトルマッチ レシーブ
	$TF{\cdot}ILF_m$	２ １ ０ 体育館 日 ３ 戦 本田 ５ ボクシング
	$TF{\cdot}IDF{\cdot}ILF_m$	体育館 博明 本田 東レ レック 久光製薬 クラティンデンジム ボクシング ＷＢＣ サク
	BIN_m	中央 体育館 大阪 市 日 １ ０ ２ ５ ３
	IDF_m	体育館 中央 ボクシング 市 大阪 バレーボール グリーンツダ 戦 回戦 クラティンデンジム
	TF_m	２ １ ０ 日 ３ ５ ４ ６ ８ 大阪
	$TF{\cdot}IDF_m$	２ １ 体育館 本田 ０ 東レ 戦 ボクシング ３ ５
国立京都 国際会館	M	京都議定書 地球温暖化防止京都会議 ＩＷＣ総会 国際捕鯨委員会
	ILF_m	左京 会館 国立 京都 国際 軍縮 会議 テロリズム カ国 明子
	$IDF{\cdot}ILF_m$	左京 軍縮 会館 エフジェニー ゴルコフスキー テロリズム 国立 京都 レオニドヴィッチ 明子
	$TF{\cdot}ILF_m$	軍縮 京都 国際 国連 会議 テロ ０ １ 会館 左京
	$TF{\cdot}IDF{\cdot}ILF_m$	軍縮 京都 国連 テロ 左京 ゴルコフスキー 会議 国際 テロリズム 会館
	BIN_m	京都 会館 国立 国際 日 市 ２ 大阪 ０ １
	IDF_m	左京 会館 国立 京都 国際 軍縮 会議 テロリズム 明子 カ国
	TF_m	京都 ０ １ 軍縮 国際 ２ 会議 テロ 国連 ５
	$TF{\cdot}IDF_m$	軍縮 京都 国際 会議 国連 テロ 会館 国立 左京 ０
トヨタ自動車	M (2 空間的文脈)	柔道 田村亮子 トヨタ 車 排ガス 自動車メーカー 車メーカー 労使交渉 企業 バスケットボール スーパーリーグ
		スピード スケート 寺尾悟 トヨタ 自動車
	ILF_m	トヨタ自動車 トヨタ 日 １ ２ ３ ０ ホンダ ４ ５
	$IDF{\cdot}ILF_m$	トヨタ自動車 トヨタ ホンダ 日産自動車 自動車 豊田自動織機 ショートトラック 小型車 雅俊 デンソー
	$TF{\cdot}ILF_m$	１ ０ ２ ３ ４ ５ トヨタ自動車 ６ ７ ８
	$TF{\cdot}IDF{\cdot}ILF_m$	トヨタ自動車 トヨタ ０ １ 秒 ２ 敗 ホンダ 位 ３
	BIN_m	トヨタ自動車 １ ２ 日 ０ ３ ５ ４ ６ ７
	IDF_m	トヨタ自動車 トヨタ ホンダ 日本 ７ ４ ５ ６ ３ ８
	TF_m	１ ０ ２ ３ ４ ５ ６ ７ ８ ９
	$TF{\cdot}IDF_m$	０ １ ２ ３ トヨタ自動車 ４ ５ ６ ７ トヨタ