社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－ＦＩ－ 93 (9)
2009－ＮＬ－189 (9)
2009／1／23

Wikipediaカテゴリ階層構造の固有名詞分類実験における効果

杉原 大悟† 増市 博†
梅基 宏† 鷹合 基行†

†富士ゼロックス株式会社 研究技術開発本部

本稿では，Wikipediaの記事タイトルを機械学習手法を用いて固有名詞クラスへと分類する際にWikipediaのカテゴリ階層を分類器の素性として用いた場合の効果について報告する．ある記事タイトルに関連するWikipediaのカテゴリネットワーク情報を分類器の素性として用いるために，あるカテゴリのカテゴリ階層構造をシンプルな基準「主要カテゴリからの最短経路」によってWikipediaのカテゴリネットワークから抽出した．Wikipediaの語を関根の拡張固有名詞階層へ機械学習手法を用いて分類する実験において，得られたカテゴリ階層構造の効果を確認した．固有名詞クラスの粒度は荒いクラス分類 (固有名詞階層の深さ 2) と細かいクラス分類 (固有名詞階層の深さ 4) の 2 種類を用意し，荒い固有名詞分類においては 7 クラスおよび 15 クラスへと分類する実験，細かい固有名詞分類においては 118 クラスへと分類する実験を行った．カテゴリ階層構造を用いない場合と比較して，カテゴリ階層構造を用いた場合に，より良い F 値が得られた．最良の分類器による精度はそれぞれ F 値 91.33，F 値 89.68，F 値 84.06 であった．また，カテゴリ階層構造は Recall の向上に特に効果的であり，その効果は分類先の固有名詞分類の粒度が細かくなれば減少することが分かった．

Effect of hierarchical category structures of Wikipedia in experiments for named entity categorization

Daigo Sugihara†, Hiroshi Masuichi†, Hiroshi Umemoto† and Motoyuki Takaai†

†Research & Technology Group, Fuji Xerox Co., Ltd.

We consider the effect of hierarchical category structures of Wikipedia in experiments for named entity categorization based on machine learning methods. We extracted hierarchical category structures of each category simply based on the shortest path from "Category:Main Category" of Wikipedia. We checked the effect of the hierarchical category structure in categorization of term of Wikpedia to named entity class defined by Sekine's Extended Named Entity Hierarchy. We prepared 3 types of experimental settings on the number and granularity of named entity classes: The first one is categorization to rough-grained 7 classes of named entity, the second is categorization to rough-grained 15 classes of named entity and the third is categorization to fine-grained 118 classes of named entity. The classifier learned with the hierarchical category structure got better than those without hierarchical category structure in F-measure. The F-measures of our best classifier are respectively 91.33%, 89.68% and 84.06% in above 3 experimental settings. Our study found that hierarchical category structures are especially effective in the recognition of named entities in categorization to rough-grained classes of named entity and its effect declines in categorization to fine-grained classes of named entity.

1. はじめに

本稿では，Wikipedia の記事タイトルを機械学習手法を用いて固有名詞クラスへと分類する際にWikipediaのカテゴリ階層を分類器の素性として用いた場合の効果について述べる．

Wikipediaは，各国のボランティアによって執筆・編集・メンテナンスが行われているWeb上の多言語電子百科事典であり，幅広い概念を網羅するコンテンツとなっている．加えて，WikipediaはWeb上のWiki形式で編集されているため，記事が記事内の語句から他の記事へのリンクによって相互参照され，記事には記事の内容に関連したカテゴリが割り振られており，また，Infoboxなどの情報抽出タスクなどに有益な情報が含まれているなど，既存の辞書にはない多

くの特性がある．そのため，Wikipediaから新たな言語資源を獲得する研究およびWikipediaと既存の言語資源を統合する研究が盛んに行われている．例えば，新たな言語資源獲得の研究に関しては，オントロジ構築[14)15)]や知識獲得[5)]，固有名詞の語義曖昧性解消[1)]や固有名詞認識[3)]などの例が挙げられる．また，語彙網羅性の高いWikipediaを他の既存の言語資源にマッピングまたは統合し，言語資源を拡充する研究としては，WikipediaのInfoboxのWordNetへの関連付け[11)]，Wikipediaの記事タイトルの日本語語彙体系への統合[13)]，WikipediaとWordNetの統合によるオントロジー構築[9)]などの例が挙げられる．特に固有名詞分類体系へのマッピングについては，本研究の先行研究として渡邉ら[16)]の研究がある．渡邉らは，箇条書きのHTML構造をグラフのクリークとして捉え，無向グラフィカルモデルであるConditional Random Fields(CRF)の適用を行っている．

これらの研究において，記事間の相互参照や記事に与えられたカテゴリなど，他の既存の言語資源にはないWikipediaの特性が利用されている．例えば，渡辺らはWikipediaの記事の相互参照性とHTMLの構造に注目している．本研究では特にWikipediaの記事に付与されたカテゴリとそのカテゴリのネットワークに注目する．Wikipediaのカテゴリは記事タイトルを分類するものや，記事タイトルに関連する語が含まれている．また，Wikipediaのカテゴリには他のカテゴリへのリンクが設定されており，カテゴリとそのサブカテゴリの間に上位下位の関係が成り立つものもある．我々は，Wikipediaのカテゴリとカテゴリのリンクの情報を適切に用いれば，語の分類が有利に行いうると考えている．

しかしながら，Wikipediaのカテゴリのリンク構造は木構造ではなく，循環を含むなど複雑な構造となっているため，タイトル語の分類のためにカテゴリのリンク構造を用いる際に，あるタイトル語に付与されたカテゴリのリンクをどの範囲まで取得すべきかは明確ではない．そして，無作為なカテゴリのリンク構造抽出は，語の分類を行う際にノイズとなる恐れがある．

そこで本研究では，あるカテゴリが持つ上位カテゴリのリンク構造をシンプルな基準「主要カテゴリからの最短経路」で明確に固定することを考える．我々は，あるカテゴリが持つ上位カテゴリリンク構造に関してこのような基準で制約することによって，あるカテゴリが持つ上位カテゴリは，主要カテゴリからの最も基本的なパスに限定され，固有名詞分類タスクにおいての分類の手がかりとして有効に活用できると考えた．以下に，この「あるカテゴリへの主要カテゴリからの最短経路として制約されたWikipediaのカテゴリネットワーク」を「カテゴリ階層構造」と呼ぶ．

本研究の研究の立ち位置と研究の目的は，以下のようにまとめられる．

- 本研究はWikipediaの記事タイトルを固有名詞に分類することを考える．
- 本研究はWikipediaのカテゴリネットワークに注目し，カテゴリが持つ上位カテゴリのリンク構造をシンプルな基準「主要カテゴリからの最短経路」で明確に固定し，カテゴリ階層構造とする．
- 本研究の目的は，カテゴリ階層構造を機械学習による固有名詞分類タスクに用いた際の効果を，実験を通して確認することである．

以下における本稿の構成について述べる．

2では，本研究での提案手法の詳細について述べる．3では，本研究で行った固有名詞分類実験について述べる．4では，関連研究について述べる．5では，まとめと今後の課題を述べる．

2. 提案手法

本研究では，カテゴリのリンク構造を固有名詞分類タスクに用いる際に，あるカテゴリが持つ上位カテゴリのリンク構造をシンプルな基準「主要カテゴリからの最短経路」で明確に固定し，Wikipediaのタイトル語を固有名詞に分類するタスクを解く際には，カテゴリ階層構造を分類器への素性として用いる．以下において，Wikipediaのカテゴリ構造を固有名詞分類タスクの手がかりとして用いる場合の問題点，本研究におけるWikipediaのカテゴリリンク構造の整理方法，本研究で得られたカテゴリ階層構造，固有名詞分類タスクにおけるカテゴリ階層構造の利用方法について述べる．

2.1 Wikipediaのカテゴリ構造の問題点

Wikipediaのカテゴリネットワークについて，Thomら[10)]によると以下の特性があるという[⋆1]．

- カテゴリには多くのサブカテゴリと上位カテゴリが存在する場合がある．
- カテゴリ間の関係はis-a等の包摂関係を持たないことがある．例えば「Category:ヨーロッパ」のサブカテゴリに「Category:NATO」が存在するが，このカテゴリ間の関係はis-a関係ではない．
- カテゴリのリンクには循環構造がある．

また，Zeschら[12)]によるグラフ理論的な分析では，Wikipediaのカテゴリグラフは，scale-freeおよびsmall-worldグラフであると述べられている．すなわち，カテゴリのリンクを芋蔓式に辿っていけば比較的簡単にWikipediaのカテゴリのどこにでもいきつくということである．そのため，タイトル語を固有名詞に分類する手がかりとするために，カテゴリのリンク構造を用いることを考えた場合，あるタイトル語に付与されたカテゴリのリンクをどの範囲まで取得すべきかは明確ではないという問題がある．カテゴリはカテ

⋆1 ここで挙げる例は，2008年12月時点の日本語Wikipediaから得たものである．

ゴリリンクを通して全てのカテゴリと繋がっているため，固有名詞分類の手がかりとして用いるためには，何らかな基準でカテゴリリンクの選択が必要である．また，カテゴリとそのサブカテゴリの関係も is-a の関係のみならず，関連する多様な関係が含まれている等，リンク間関係にも規則性に乏しい面がある．そのため，記事に付与されたカテゴリからボトムアップに特定の関係 (例えば is-a の上位カテゴリのみ) のカテゴリを取得して，語を分類する際に手がかりとすることを考えたとしても，関係の認定のための知識源が必要であり，また，取得の範囲を適切に指定しなければならない．

2.2 カテゴリリンク構造の整理方法

そこで，本研究では，あるカテゴリが持つ上位カテゴリのリンク構造をシンプルな基準「主要カテゴリからの最短経路」として明確に固定することを考える．このカテゴリ階層構造を固有名詞分類タスクに用いた際に考えられるメリットを以下に挙げる．

- 主要カテゴリからの経路にカテゴリの上位語が含まれることが期待され，固有名詞分類の際のマーカーとして働くことが期待される．
- 主要カテゴリからトップダウンに各カテゴリへの経路を設定するため，固有名詞ごとに共通の上位カテゴリが定まり，固有名詞分類の際のマーカーとして働くことが期待される．
- カテゴリリンク関係 (is-a 等) の選択などのために Wikipedia の外部に知識源を必要としない．

本研究における主要カテゴリから各カテゴリへの最短経路の取得の手順は以下のようになる．

STEP1　以下の STEP2 を C_0(主要カテゴリ) からスタートし，各カテゴリに対する最短経路候補を取得する．

STEP2　C_i はカテゴリ i を表し，CS_j は任意のカテゴリの集合とし，$Route\_A = [C_0, CS_1, CS_2, ..., C_A]$ を C_A に至る経路とする (経路はリストとして表現してあり，リストの各要素はリスト内で次に続く要素へのパスを持つとする)．カテゴリ C_A がサブカテゴリを持ち，かつ，経路の長さが L(今回は $L = 10$ とした) 未満の場合，C_A が持つ全てのサブカテゴリに対して以下の処理を行う．

2-1 C_A のサブカテゴリ $C_{A'}$ に対して経路 $Route\_A' = [C_0, CS_1, CS_2, ..., C_A, C_{A'}]$ をサブカテゴリ $C_{A'}$ の最短経路候補とする．

2-2 サブカテゴリ $C_{A'}$ に対する最短経路候補がすでに他のカテゴリ C_B のサブカテゴリとして得られている，あるいはカテゴリ集合 CS_B からの経路として得られている場合には，$C_{A'}$ への経路に関して集約を行う．$CS_C = [C_A, C_B]$，または，$CS_C = [C_A | CS_B]$ とし，経路 $Route\_A' = [C_0, CS_1, CS_2, ..., CS_C, C_{A'}]$ とする．

STEP3　各カテゴリの最短経路候補のうち実際に最も短い経路を各カテゴリの最短経路とする．

また，今回の最短経路の抽出では，上記の通り最大経路数 L を 10 とした．各カテゴリへの主要カテゴリからの最短経路を計算する際に最大経路数を 11 以上に設定しても，最短経路が得られたカテゴリ数がさほど増加しなかったことから，計算時間との兼ね合いで 10 と決定した．

2.3 主要カテゴリからの最短経路例

上記の手順で得られたカテゴリの主要カテゴリからの最短経路例を表 1 に示す．表では，記号">"を用いてカテゴリ間のリンク階層を,「上位カテゴリ > サブカテゴリ」のように表す．また，表記の簡単のために，共通のサブカテゴリまたは上位カテゴリを含み主要カテゴリからの同一距離のカテゴリ同士については,[A,B] のようにまとめて表示している．但し，カテゴリをまとめて表す場合には，">"記号の左右のカテゴリ集合の全ての組み合わせにカテゴリリンクがあるわけではないことに注意されたい．例えば,「[社会, 歴史]>[テーマ史, 政治]」の表示部分では,「歴史」と「テーマ史」の間にはカテゴリリンクがあり,「社会」と「政治」の間にはカテゴリリンクがある．これは「テーマ史」と「政治」には共通のサブカテゴリ「政治史」が存在するためにまとめて表示され,「テーマ史」と「政治」の上位のカテゴリである「社会」と「政治」は共通の上位カテゴリ「主要カテゴリ」を持つため，これらのカテゴリもまとめて表示した結果である.「[政治史, 日本のテーマ史]> 日本の政治史」のように">"の左右どちらかが単体のカテゴリの場合は，左右のカテゴリ集合中のカテゴリとカテゴリの間にはリンクが存在する．

表 1 において,「Category:横須賀市」と「Category:横須賀市出身の人物」の上位カテゴリ階層構造における共通の部分は主要カテゴリのみとなり，双方のカテゴリ間に意味的な曖昧性が生じていないことが分かる．逆に,「Category:横須賀市の鉄道駅」への経路には，上位カテゴリに「Category:地理」を持つ経路および上位カテゴリに「Category:社会」を持つ経路の 2 つが得られた.「Category:日本の内閣総理大臣」に到る経路においても,「Category:日本の政治史」より上の階層で複数の経路が存在する．このように，各カテゴリの上位カテゴリを主要カテゴリからの最短経路で選択した場合にも，各カテゴリへの経路は複数生じる場合があるが，今回の固有名詞分類実験においては最短ならば経路を複数採用することとした．

2.4 カテゴリ階層構造の利用方法

本研究では，Wikipedia のタイトル語を固有名詞クラスへと分類する際に，カテゴリ階層構造を以下の手順によって分類器への素性として与えた．

- タイトル語の記事からカテゴリを得る．

表 1　主要カテゴリからの最短経路例

カテゴリ名	主要カテゴリからの最短経路
横須賀市出身の人物	主要カテゴリ ＞ 人間 ＞ 人物 ＞ 出身地別の人物 ＞ 日本出身の人物 ＞ 神奈川県出身の人物 ＞ 横須賀市出身の人物
日本の内閣総理大臣	主要カテゴリ ＞[社会, 歴史]＞[テーマ史, 政治]＞[政治史, 日本のテーマ史]＞ 日本の政治史 ＞ 日本の内閣総理大臣
横須賀市	主要カテゴリ ＞ 地理 ＞ 地方行政区画 ＞ 各国の地方区分 ＞ 日本の地方行政区画 ＞ 日本の市町村 ＞ 神奈川県の市町村 ＞ 横須賀市
横須賀市の鉄道駅	主要カテゴリ ＞ 地理 ＞ 地方行政区画 ＞ 各国の地方区分 ＞ 日本の地方行政区画 ＞ 日本の市町村 ＞ 神奈川県の市町村 ＞ 横須賀市 ＞ 横須賀市の鉄道駅 主要カテゴリ ＞ 社会 ＞ 各国の社会 ＞ 日本の社会 ＞ 日本の交通 ＞ 関東地方の交通 ＞ 関東地方の鉄道駅 ＞ 神奈川県の鉄道駅 ＞ 横須賀市の鉄道駅

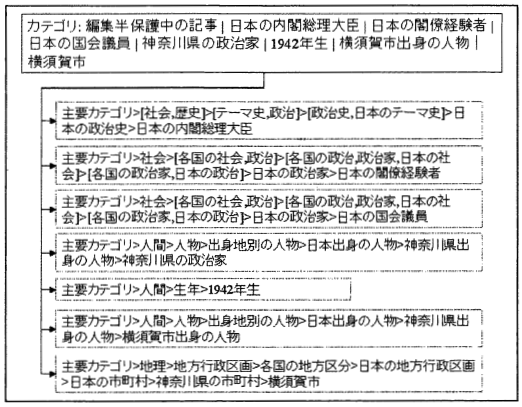

図 1　カテゴリの拡張

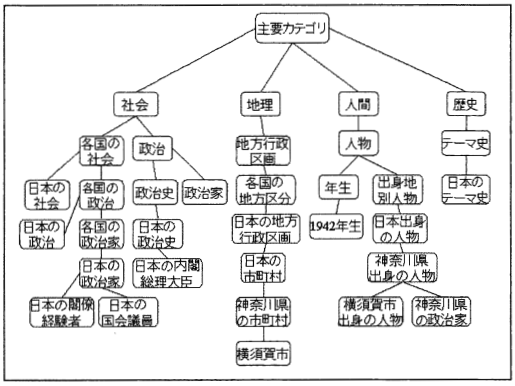

図 2　カテゴリ階層構造から構築した木

- カテゴリをカテゴリ階層構造へと拡張する．
- タイトル語のカテゴリの階層構造から主要カテゴリを根とする 1 つの木を構築する．
- 木から実験設定ごとの素性を生成する．

図 1 では，記事「小泉純一郎」のカテゴリとそのカテゴリ階層構造への拡張過程を示している．その際，「Category:編集中半保護中の記事」など Wikipedia の管理用のカテゴリは除去している．また，図 2 では，これらの拡張されたカテゴリの経路を「主要カテゴリ」を根とする木に変換している．木に変換したのは，今回の固有名詞分類実験において，得られたカテゴリ階層構造中のカテゴリの頻度の影響を除き，上位カテゴリの利用によって固有名詞分類タスクの精度がどのようになるかを観察するためである．

3. 固有名詞分類実験

本研究では Wikipedia カテゴリの階層構造を機械学習による固有名詞分類タスクに用いた際の影響を確認するため，Wikipedia のタイトル語を固有名詞クラスへ分類する実験を行った．固有名詞体系には，先行研究である渡邉ら[16] に習い，関根ら[7] による拡張固有名詞階層の分類体系を用いた．情報検索などでは粒度の細かい固有名詞分類体系が望ましい場合があることから，拡張固有階層の深さ 2 層までのクラスを用いた場合と，深さ 4 層までのクラスを用いた場合の 2 種類で実験データを作成した．固有名詞階層 2 階層 (以下，NE2 層と呼ぶ) については 7 クラスと 15 クラスの分類実験，固有名詞階層 4 階層 (以下，NE4 層と呼ぶ) は 118 クラスの分類実験を行った．分類器には SVM と工藤ら[4] による木の分類システムを用いた．以下に，実験データ，実験方法，および実験結果と考察について述べる．

3.1 実験データ

3.1.1 主要カテゴリからの最短経路取得結果

日本語 Wikipedia の 2008 年 4 月 8 日の dump ファイルから 48830 のカテゴリを取得し，Wikipedia のカテゴリのうち管理用のカテゴリを除外するための表 3 に示すフィルターを通過した 45231 カテゴリを最短経路取得の対象カテゴリとした．対象カテゴリのうち最短経路が得られたカテゴリ数は 36598 であった．よって，対象としたカテゴリのうち最短経路が抽出できた割合は 80.91% であった．また，最短経路の抽出に失敗したカテゴリを分析したところ，「1977 年生」や「1977 年没」などの生年/生没に関するカテゴリが 3637 あることが分かった．これらのカテゴリには，最短経路が取得できた生年/生没のカテゴリの経路を参考に人手で経路を与えることとした．結果，40235 のカテゴリに経路を与えることができ，対象カテゴリのうち 88.95% が上位カテゴリの情報を持つことになった．これらの数値を表 4 に示す．

表 2 固有名詞クラスの内訳

固有名詞クラス（深さ 2 以上）	固有名詞階層 3 層以下の例	クラス数	語数
名前-地名	GPE-市区町村名, 地形名-陸上地形名, 地形名-河川湖沼名, GPE-国名, 地形名-海洋名,GPE-群名, 地名_その他　など	16	6261
名前-施設名	路線名-道路名,GOE-学校名,GOE-神社寺名,GOE-駅名-電車駅名 GOE-GOE_その他, 路線名-電車路線名,GOE-公共機関名　など	22	5690
名前-製品名	芸術名-番組名, 芸術名-文学名, 製品名_その他, 芸術名-映画名, 規則名, 言語名, 乗り物名-車名, 医薬品名, 芸術名-音楽名　など	34	5407
名前-組織名	企業名, 組織名_その他, 協会名, 政府組織名, 国際組織名 競技グループ名, 民族名, 政党名, 軍隊名, 企業グループ名　など	11	4988
名前-人名	階層なし	1	4148
名前-自然物名	物質名, 生物名-動物名-脊椎動物-鳥類, 生物名-動物部位名 生物名-動物名-脊椎動物-哺乳類　など	13	2364
名前-イベント名	戦争名, 催し物名-催し物名_その他, イベント名_その他 催し物名-大会名, 事故事件名, 自然災害名, 催し物名-会議名	7	945
時間表現・数値	日付表現, 時代表現, 曜日表現, 数値表現-個数	4	938
名前-病気名	階層なし	1	585
名前-称号名	階層なし	1	404
名前-職業名	階層なし	1	191
名前-神名	階層なし	1	145
名前-色	階層なし	1	76
名前-単位名	単位名_その他, 通貨名	2	53
名前-名前_その他	名前-名前_その他, 名前-識別番号 など	3	46
TOTAL		118	32241

表 3 管理用カテゴリを除外するフィルター条件

カテゴリの文字列に関する条件	先頭が「User_」
	末尾が「の画像」
	末尾が「ユーザー」
	末尾が「スタブ項目」
	末尾が「のスタブ」
	末尾が「のテンプレート」
	末尾が「ウィキペディアン」

表 4 カテゴリの経路の抽出結果数

	異なり数	対象カテゴリ比率 (%)
全カテゴリ	48830	107.96
対象カテゴリ	45231	100.00
最短経路抽出に成功 (*1)	36598	80.91
最短経路抽出に失敗 (*2)	8633	19.09
*2 の生年/生没のもの (*3)	3637	8.04
経路付与カテゴリ (*1 と*3)	40235	88.95

3.1.2 学習および評価データ

学習および評価用データには，日本語版 Wikipedia(2008 年 4 月 8 日の dump) の記事タイトルに対して関根ら[7] による拡張固有名詞階層の分類を付与したものを用いた．具体的には，渡邉らによる「NAIST Japanese ENE Dictionary on Wikipedia(NAIST-jene) ⋆1」のデータのうち日本語版 Wikipedia(2008 年 4 月 8 日の dump) に記事が存在しカテゴリを持つ語，および，関根らが公開している固有名詞分類データ ⋆2のうち日本語版 Wikipedia(2008 年 4 月 8 日の dump) に記事が存在しカテゴリを持つ語を用いた．これら 2 つのリソースを同時に用いた理由を以下に列挙する．

- 双方ともに関根の拡張固有名詞階層の Ver6.1.4 の定義によるデータであり，一貫性が保証されている．
- 多クラス分類実験となるため，なるべくデータが多いほうが望ましい．

NE2 層までの固有名詞データでは，渡辺らと同様に数が極端に少ない固有名詞クラス (頻度 25 未満) については，1 つのクラス「名詞-名前_その他」に統合した．時間および数量表現も 1 つのクラスに統合した．結果 NE2 層までのクラス数は 15 となった．NE4 層の固有名詞データでは，語の頻度 5 以上のものを選択し，クラス数は 118 となった．固有名詞の内訳を表 2 に示す．NE2 層の頻度上位 7 クラスのデータを用いた 7 クラス分類実験，NE2 層の全 15 クラスのデータを用いた 15 クラス分類実験，NE4 層の全 118 クラスを用いた 118 クラス分類実験を行った．また，渡邉らは HTML の箇条書きの構造を分類システムにとりこみ，Wikipedia に記事のない語の分類を行っているが，本研究ではカテゴリ階層構造を扱っており，Wikipedia に記事のない語を分類対象とはしないため，渡邉らの研究と本研究の精度の比較については考察しない ⋆3．

⋆1 http://cl.naist.jp/ masayu-a/p/NAIST-jene.html
⋆2 http://nlp.cs.nyu.edu/ene/

⋆3 渡邉らは，箇条書き中の語を元記事へのリンクのない語も含めて固有名詞 13 クラスへ分類する実験を行い，F 値 78.62 の精度を報告している．本研究による 15 クラス分類実験の F 値は 89.68 であるが，分類対象の記事タイトル語にはカテゴリを持つ語を選ぶなど有利な条件であり，比較はできない．

分類対象のタイトル語「エディ・オフォード」(名前-人名)
・カテゴリ階層構造なしの木のS式(主辞) (主要カテゴリ(音楽プロデューサー)(イエス)) →「名前-職業名」へ分類
・カテゴリ階層構造ありの木のS式(表層) (主要カテゴリ(人間(人物(各国の人物(イギリスの人物(イギリスの音楽プロデューサー)))))(文化(サブカルチャー(音楽のムーブメント(ロック(ロック・バンド(イエス)))))(娯楽(プロデューサー(音楽プロデューサー))))) →「名前-人名」へ分類

図 3 カテゴリ階層構造を用いた場合に正しく分類した例

分類対象のタイトル語「枢機卿」(名前-称号名)
・カテゴリ階層構造なしの木のS式(主辞) (主要カテゴリ(称号)(カトリック)(枢機卿)(バチカン)) →「名前-称号名」へ分類
・カテゴリ階層構造ありの木のS式(表層) (主要カテゴリ(社会(国(大陸別の国(ヨーロッパの国(バチカン))))(思想(宗教(キリスト教(キリスト教の教派(カトリック))))(生活))(人間(人の一生)(人物(宗教に関連する人物(キリスト教に関連する人物(枢機卿)))(称号)))(文化(地域別の文化(西洋文化))) →「名前-製品名」へ分類

図 4 カテゴリ階層構造を用いた場合に誤って分類した例

3.2 実験方法

分類システムには，SVM の実装として TinySVM を [*1]，木の分類システムの実装には BACT [*2]をそれぞれ用いた．SVM のカーネルは予備実験の結果から線形カーネルを用いた．TinySVM も BACT も 2 値分類器であり，多クラス問題への適用のため，one-versus-rest 法を用いた．語に対する固有名詞の決定には，クラスの数だけの 2 値分類器で「その語が特定の固有名詞か否か」の分類を行い，分類器の出力する値が閾値 (=0.0) を越える場合，分類器の出力値が最も高い固有名詞クラスに分類した．その学習と評価は上記データの 5 分割交差検定にて行った．

実験設定には以下の基準から 21 種類の設定を用意した．

- 分類クラス数が (7 | 15 | 118) クラス．
- 分類システムが (TinySVM | BACT)．
- カテゴリ階層を (用いない | 用いる)．
- カテゴリの文字列の (主辞のみ | 主辞と表層文字列) を用いる．

但し，カテゴリ階層構造を用いた場合の BACT による実験では，あるカテゴリの主辞をカテゴリ階層構造による木のどこに配置してよいか明確ではないため表層のみを用いることにした．

BACT での学習および評価の際には，カテゴリ階層構造の木を S 式で表現し，システムへの入力としている．TinySVM には，この木のノードのカテゴリの文字列から主辞と表層を取り出し，「そのカテゴリの文字列があるかないかを表す素性」の値を 1 としてベクトルを生成する．つまり，SVM の分類器への素性にカテゴリの頻度や木の構造に関するものは設定されていない．

3.3 実験結果と考察

実験結果を表 5 に示す．カテゴリ階層構造を用いた実験結果は，用いない場合に比べて Recall が高く，Precision が若干劣る結果となっている．これは，特に固有名詞の粒度の荒い NE2 層 7 クラスと NE2 層 15 クラスの分類実験結果に顕著である．NE2 層の実験結果においては，カテゴリ階層構造を用いた場合とカテゴリ階層構造を用いない場合を比べて，SVM の実験設定では 2 から 4 ポイント程度 Recall がよく，BACT の実験設定では 5 から 6 ポイント程度 Recall がよい．逆に Precision においては，カテゴリ階層構造を用いた SVM の実験設定では 1 から 2 ポイント程度悪く，BACT の実験設定では 2 ポイントから 3 ポイント程度悪くなっている．カテゴリ階層構造を用いることで，Recall の向上が Precision の悪化より勝った結果，F 値において精度の向上を得た．

図 3 に NE2 層 15 クラス分類実験において，カテゴリ階層構造を用いた分類器が正しく分類でき，カテゴリ階層構造を用いない場合に誤った分類を行う例を挙げる．分類対象のタイトル語は「エディ・オフォード」という「名前-人名」であるが，カテゴリ階層構造を用いない場合には「名前-職業」へと誤って分類してしまっている．カテゴリ階層構造を用いない場合はカテゴリの表層である「音楽プロデューサー」という職業名が影響するためだと考えられる．対して，カテゴリ階層構造を用いる場合には，上位のカテゴリに「Category:人物」などが存在するため「名前-人名」へと正しく分類できたものと考えられる．このような事例によって，カテゴリ階層構造を用いることで固有名詞分類タスクにおいて Recall が向上するものと考えられる．逆に，図 4 には，NE2 層 15 クラス分類実験において，カテゴリ階層構造を用いた場合にのみ分類器が誤った分類を行う例を挙げる．この場合，カテゴリの表層に「称号」があるため，カテゴリの表層のみで正しく「名前-称号名」へと分類が可能となると考えられる．対して，カテゴリ階層構造には「Category:思想」などの他の固有名詞クラスと関連が高い上位カテゴリが含まれてしまい，結果としてカテゴリ階層構造を用いた分類器は「名前-製品名」へ誤って分類したものと考えられる．このような事例のために，カテゴ

*1 http://chasen.org/ taku/software/TinySVM/
*2 http://chasen.org/ taku/software/bact/

表 5 実験結果

実験設定	NE2 層 7 クラス R	P	F	NE2 層 15 クラス R	P	F	NE4 層 118 クラス R	P	F
SVM _C 階層なし (主辞)	85.63	94.15	89.69	84.01	93.53	88.50	77.38	**90.92**	83.60
SVM _C 階層なし (主辞+表層)	85.43	93.82	89.43	84.43	93.17	88.57	78.35	90.43	83.95
SVM _C 階層あり (主辞)	**89.67**	93.06	**91.33**	**87.14**	92.38	**89.68**	79.59	89.09	**84.06**
SVM _C 階層あり (主辞+表層)	89.65	92.51	91.06	87.13	91.96	89.48	**79.83**	87.77	83.61
BACT _C 階層なし (主辞)	82.42	94.79	88.17	79.84	93.95	86.29	77.09	88.35	82.33
BACT _C 階層なし (主辞+表層)	81.81	**95.35**	88.06	79.49	**94.44**	86.28	77.04	88.76	82.48
BACT _C 階層あり (表層)	88.30	92.70	90.44	85.51	92.07	88.66	78.22	87.64	82.65

注意)「C 階層」とは「カテゴリ階層」を表す.

リ階層構造を用いることで Precision が悪化したものと思われる.

また, 固有名詞の粒度が細かい NE4 層 118 クラスのデータによる分類実験では, カテゴリ階層を用いた場合の精度の向上が, F 値において SVM では 0.1 ポイント, BACT では 0.2 から 0.3 ポイント程度にとどまり, NE2 層の実験設定に比べて小さな値になっている. 分類先の固有名詞の粒度が細かい場合には, ある固有名詞を特徴的に表すカテゴリを記事が持つかどうかが重要になるのではないかと考えられる. また, 素性を作成する際にカテゴリの文字列を主辞のみ用いる場合と表層文字列を含める場合の精度の差は小さなものになっており, 場合によっては表層を含めることで精度の悪化を招くようである. また, 7 クラス, 15 クラス, 118 クラスの分類実験全てにおいて, 差の大小はあるが, 最も F 値がよいのは「SVM _カテゴリ階層あり (主辞)」であり, それぞれ F 値 91.33, F 値 89.68, F 値 84.06 となった. この結果により, Wikipedia のカテゴリ階層は固有名詞分類タスクに有効であることが示されたと考える.

4. 関 連 研 究

以下に, 本研究の関連研究について述べる. 特に, Wikipedia を既存の言語資源にマッピングし言語資源を拡充する研究, および, Wikipedia のカテゴリに関する研究について, それぞれ本研究との関連を述べる.

4.1 既存言語資源へのマッピングに関する関連研究

渡邉ら[16] は Wikipedia から得た語を固有名詞に分類する研究を行っており, 同様のタスクを解く本研究の先行研究である. 渡邉らのモデルは, 箇条書きの HTML 構造をグラフのクリークとして捉えた CRF によるモデルであり, 箇条書き中の語の分類結果が他の語を分類する際にも波及する仕組みであり, 元記事へのリンクがない場合にも箇条書き中の他の記事を分類の手がかりとして用いることができる. 本研究との比較であるが, 我々の語の分類システムはカテゴリのリンク構造に注目し, 渡辺らは Wikipedia の記事の相互参照性と HTML の構造に注目している点で志向性が異なる. 我々の語の分類システムは, 記事に振られたカテゴリのリンク構造を利用するため, 記事へのリンクのない語を分類することはできないが, 渡辺らのシステムはカテゴリのリンク構造は考慮していない. 我々の研究成果を渡邉らのシステムに容易に取り込むことができるため, 双方の研究は相補的な関係にあると言える.

4.2 Wikipedia のカテゴリについての関連研究

Wikipedia の記事やカテゴリのリンク構造を利用する関連研究には, 以下の研究がある.

Schonhofen[6] はカテゴリを文書分類におけるトピック特定に利用した研究を行っている. ただし, Schonhofen はカテゴリのリンク構造を利用していない.

桜井ら[14] はカテゴリとサブカテゴリの文字列の文字列照合部分から is-a の認定を行う手法を日本語オントロジ構築に適用している. 桜井らの目的はオントロジーの構築であるが, Wikipedia のカテゴリ構造を固有名詞分類に用いる場合にも, 文字列の照合等によってカテゴリ間の関係 (is-a 等) を認定し, その情報をカテゴリ階層構造と共に分類器の学習素性に用いるなどの応用が考えられる.

単語の関連度の算出については, Strube らによる WikiRelate[8] の研究がある. WikiRelate では, Wikipedia のカテゴリネットワーク上の距離や共通カテゴリの数などを基に算出する. 関連度の算出に関して, Gabrilovich ら[2] は記事から引いた単語群から重み付ベクトルを生成し, ベクトルのコサイン尺度等によって関連度を算出する手法を提案しており, 実験による精度が WikiRelate よりも高かったことを報告している. これについて中山ら[15] はカテゴリの距離は関連度算出のパラメータとしては粗すぎること, カテゴリ間の関係は包摂的な関係以外の関係が混在していること, カテゴリのリンクは記事間の相互参照に比べて情報量が少ないことなどから関連度の算出には不適であると指摘している. 本研究ではカテゴリのリンク構造を語と語の関連度の算出には用いず, 語のクラスへの分類のための分類器の素性として用いた. 本研究において, カテゴリのリンク構造は, ある語が持つカテゴリ階層構造中にどのようなカテゴリが含まれるか, あるカテゴリの上位カテゴリとして何が存在するか, 等の ON-OFF の表現として用いられる. ただし, 今

回の実験はカテゴリ関係の混在の語分類タスクへの影響を論じる構成とはなっていない．また，Wikipediaのカテゴリが語の数に比べて十分な量か否かについては今後検討すべきであると考える．

5. まとめと今後の課題

本研究では，Wikipediaの記事タイトルを機械学習手法を用いて固有名詞へと分類する際にWikipediaのカテゴリ階層を分類器の素性として用いた場合の効果について考察し，Wikipediaのカテゴリ階層構造は固有名詞分類タスクに有効であることを確認した．本研究では，Wikipediaのカテゴリ階層を分類器の素性として用いるために，あるカテゴリが持つ上位構造をシンプルな基準「主要カテゴリからの最短経路」で明確に固定することを提案した．Wikipediaのカテゴリ階層構造を分類器の素性として用いる場合の効果について得られた知見は以下に列挙できる．

- カテゴリ階層構造は固有名詞分類タスクのRecall向上に特に有効である．
- カテゴリ階層構造の固有名詞分類における効果は分類先の固有名詞の粒度に影響を受け，固有名詞の粒度が細かくなれば効果が減少する．

今後の課題としては以下を考えている．

- Wikipediaのタイトル語の固有名詞以外の言語資源 (例えばEDR) へのマッピング．
- Wikipediaのカテゴリ階層とマッピング先の意味体系の粒度に関する考察．
- カテゴリ間の関係 (is-a) 等の利用．
- 一覧のカテゴリや記事タイトルの定義文のカテゴリの利用．

参 考 文 献

1) Razvan Bunescu and Marius Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06), 2006.*
2) Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI 2007), pp.1606-1611, 2007.*
3) Jun'ichi Kazama and Kentaro Torisawa. Exploiting Wikipedia as External Knowledge for Named Entity Recognition. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007), pp. 698-707, June, 2007.*
4) Taku Kudo and Yuji Matsumoto. A Boosting Algorithm for Classification of Semi-Structured Text. *EMNLP, 2004 .*
5) Vivi Nastase and Michael Strube. Decoding Wikipedia Categories for Knowledge Acquisition. *AAAI '08, pp.1219- 1224, 2008.*
6) Peter Schonhofen. Identifying document topic using the Wikipedia category network. *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), 2006.*
7) Satoshi Sekine, Kiyoshi Sudo, Chikashi Nobata. Extended Named Entity Hierarchy. *Proceedings of the LREC 2002, 2002.*
8) Michael Strube and Simone Paolo Ponzetto. WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI '06, pp. 1419-1424, 2006.*
9) Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum Yago: a core of semantic knowledge. *Proceedings of 16th international conference on World Wide Web, ACM, pp.697-706, 2007.*
10) James A. Thom, Jovan Pehcevski, Anne-Marie Vercoustre. Use of Wikipedia Cateogories in Entity Ranking. *Proceedings of 12th Australasian Document Computing Symposium, Melbourne, Australia, December 10, 2007.*
11) Fei Wu and Daniel S. Weld. Automatically Refining the Wikipedia Infobox Ontology. Proceeding of the 17th international conference on World Wide Web , Beijing, China, pp. 635-644, April, 2008.
12) Torsten Zesch and Iryna Gurevych. Analysis of the Wikipedia Category Graph for NLP Applications. Proceedings of the TextGraphs-2 Workshop (NAACL-HLT), Rochester, April, 2007.
13) 小林暁雄, 増山繁, 関根聡. 日本語語彙体系と日本語ウィキペディアにおける知識の自動結合による汎用オントロジー構築手法. 情報処理学会研究報告 2008-NL-187, 2008.
14) 桜井 慎弥, 手島 拓也, 石川 雅之, 森田 武史, 和泉 憲明, 山口 高平. 汎用オントロジー構築における日本語 Wikipedia の適用可能性. 人工知能学会第 18 回セマンティック Web とオントロジー研究会 (SIG-SWO-A801-06), 2008.7.
15) 中山浩太郎, 原隆浩, 西尾章治郎. 自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築. 日本データベース学会論文誌, Vol.7, No.1, pp.67-72, 2008.6.
16) 渡邉陽太郎, 浅原正幸, 松本裕治. グラフ構造を持つ条件付確率場による Wikipedia 文書中の固有表現分類. 人工知能学会論文誌, Vol.23, No.4, 2008.