Wikipediaの企業名記事における国属性の構造化

東龍太郎 ‡† 吉田光男 ‡‡† 梅村恭司 ‡‡‡†

† 豊橋技術科学大学 〒 441–8580 愛知県豊橋市天伯町雲雀ヶ丘 1–1
E-mail: ‡ r163302@edu.tut.ac.jp, ‡‡ yoshida@cs.tut.ac.jp, ‡‡‡ umemura@tut.jp

あらまし DBpedia や Yago など，Wikipedia を用いた知識ベースの構造化プロジェクトが盛んに行われている．しかし，これらは構造化された属性の数や属性値が Wikipedia 記事毎に異なるといった体系的な面で問題がある．森羅プロジェクトは Wikipedia の体系的な構造化を目的としたプロジェクトである．本稿では，森羅プロジェクトにおいて関根の拡張固有表現辞書で企業名に分類された Wikipedia 記事について，Wikipedia 記事本文とカテゴリから国に関する属性の構造化手法を提案する． 提案手法とベースラインを組み合わせることで，ベースラインと比較して再現率，F 値が向上したことを報告する．

キーワード Wikipedia，構造化，属性値抽出

1 はじめに

Wikipedia は日本語版において 100 万記事以上と大規模で，かつアクセスが容易な固有表現に関する情報源として言語処理の分野では盛んに用いられている．Wikipedia はやカテゴリなどといった半構造化された部分もあるが，記事中でも多くの情報を持つ記事本文は自然言語により記述されているため，計算機が理解しやすい形式にするためには構造化する必要がある．

2 章でも述べるが，既存の Wikipedia 構造化プロジェクトは，属性値抽出に対して一貫性の問題を抱えている．森羅プロジェクトは属性定義により一貫性の問題を解決するプロジェクトで，現在進行中のものである．

本稿では，森羅プロジェクトにおいて拡張固有表現カテゴリ「企業名」に分類された Wikipedia 記事について，Wikipedia カテゴリと本文の情報を国に関する属性の構造化手法を提案し，提案手法とベースラインを組み合わせることで，ベースラインと比較して再現率，F 値が向上したことを報告する．

2 関連研究

DBpedia [1] は Wikipedia 記事にある Infobox，カテゴリなどから構造化された知識ベースを作成する．これには，Infobox の値をそのまま用いているため，属性の数や属性名が違うといった，構造化における属性定義の一貫性がないという問題がある．例えば，「トヨタ自動車」と「日産自動車」の Wikipedia 記事は同じ Infobox クラス「基礎情報　会社」を持つが，「トヨタ自動車」は 22 の属性を持つのに対し，「日産自動車」は 44 の属性を持っており，属性の数に一貫性がない．

Yago [2] は Wikipedia 記事を WordNet [3] のノードにマッピングし構造化された知識ベースを作成する．属性は DBpedia と同様 Infobox を用いているため，一貫性の問題がある．

森羅プロジェクト [4–6] は既存の Wikipedia 構造化プロジェクトの問題点である属性の一貫性の問題を解決したものである．Wikipedia 記事を拡張固有表現カテゴリに分類 [7] し，分類された記事について構造化を行う．その際，属性数などの定義がカテゴリごとにされている．表 1 は，拡張固有表現カテゴリ「企業名」に分類された Wikipedia 記事「スターバックス」の構造化の一部である．「企業名」の属性数は 33 と固定であり，一貫性の問題を解決している．現在進行中のプロジェクト段階では，「人名」，「企業名」，「空港名」，「市区町村名」，「化合物名」の 5 つのカテゴリに分類された記事の構造化を目指している．

本稿では，森羅プロジェクトのうち，拡張固有表現カテゴリ「企業名」に分類された Wikipedia 記事について，「本拠地国」，「創業国」に着目した構造化手法とその評価結果を報告する．「本拠地国」，「創業国」に着目した理由は，正解の属性値が他の記事の正解からなる集合の要素の一つになるため，この集合に対して Wikipedia 記事を分類する文書分類問題として考えられるためである．表 1 のように Wikipedia 記事「スターバックス」において，事業内容や代表者などは Wikipedia 記事ごとに違う記述がされるが，国名について記述される「本拠地国」，「創業国」は，同じ属性値が他の記事でも出現する場合がある．

表 1　スターバックスの構造化情報 (一部)

属性名	属性値
事業内容	コーヒーストアの経営，コーヒー及び関連商品の販売
業界	飲食業
本拠地国	アメリカ合衆国
創業国	アメリカ合衆国
売上高（単体）	107 億米ドル
代表者	ハワード・シュルツ，マーティン・コールス

3 提案手法

3.1 Wikipedia 記事本文からの属性値抽出

Wikipedia 記事のタイトルについて，そのタイトルの属性値は Wikipedia 本文から以下の 2 通りで抽出できると仮定した．

（1） 本文中で最初に出現したものを抽出

手法の概要図を図 1 に示す．この手法は学習データから得られる属性値リストの中でも，本文中で最初に出現したものが Wikipedia 記事のタイトルについての属性値と推定できるという仮定に基づいている．図 1 の Wikipedia 記事「スターバックス」では，属性値リストにある属性値の中でも,「アメリカ合衆国」が Wikipedia 記事本文で最初に出現していることから,「アメリカ合衆国」を属性値としている．

（2） 本文中に最も多く出現したものを抽出
この手法の概要図を図 2 に示す．この手法は学習データから得られる属性値リストの中でも，本文中に最も多く出現したものが (1) と同じく，Wikipedia 記事のタイトルについての属性値と推定できるという仮定に基づいている．図 2 の Wikipedia 記事「スターバックス」では，属性値リストにある属性値の中でも,「日本」が Wikipedia 記事本文で最も多く出現していることから,「日本」を属性値としている．

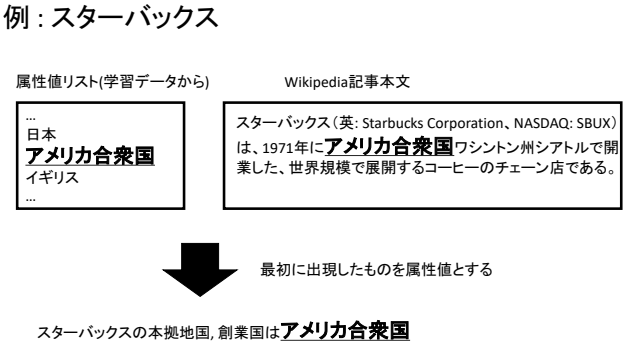

図 1　手法 (1) の概要図

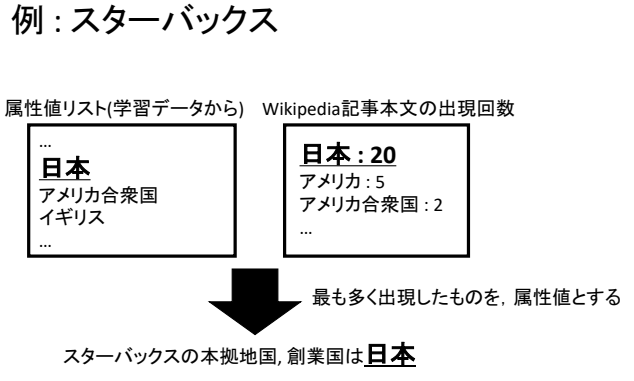

図 2　手法 (2) の概要図

3.2　Wikipedia カテゴリからの属性値抽出

手法の概要図を図 3 に示す．4.1 節でも実験データについて述べるが，Wikipedia 記事は日本語版のものであり，その中でも拡張固有表現カテゴリ「企業名」に該当する記事から「本拠地国」と「創業国」属性という国に関する属性値を抽出する．日本語版 Wikipedia における企業名の記事では，記事タイトルの企業について，その企業が日本の企業であれば，国について本文中では「日本」と述べず，都道府県等で場所を示す場合が多いことが考えられる．そこで，国に関する属性については，その記事の Wikipedia カテゴリ内に国の表記があることを仮定し，そのカテゴリから抽出する．

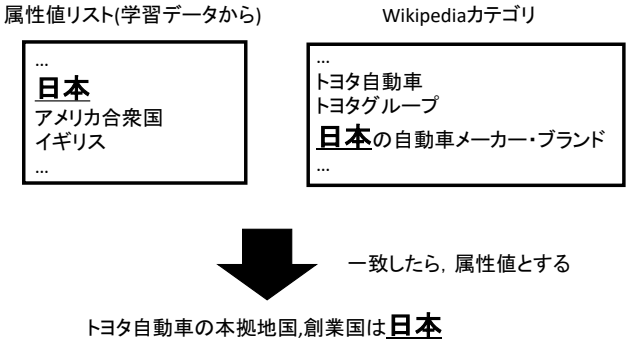

図 3　Wikipedia カテゴリを用いた手法の概要図

3.3　記事本文とカテゴリの両方からの属性値抽出

この手法は 3.1 節と 3.2 節の手法を組み合わせたものである．組み合わせるにあたり，どの手法で抽出できた属性値を優先するかが問題となる．このとき，属性値が抽出できなかった場合にのみ別の手法を用いる．

予備実験では 3.1 節の (1) と (2) の手法を用いて属性値抽出を行った．4.4 節のにおいて，3.1 節の (1) と (2) の結果を比較すると，適合率，再現率，F 値の全てにおいて (1) の手法が高かったため，3.2 節の手法と組み合わせる手法は (1) とする．

4　評 価 実 験

4.1　実験データ

実験に用いるデータについて説明する．まず，日本語版 Wikipedia のデータは 2017 年 11 月 6 日付けの CirrusSearch のダンプデータ [1]を用いる．そのうち，属性値抽出をするデータは拡張固有表現カテゴリ「企業名」に該当する Wikipedia 記事 600 件，そのうち学習データを 540 件，テストデータを 60 件とし，10 分割交差検証をする．また，正解とする属性値のデータは人手によりアノテートしたものを用いる．

4.2　ベースライン

ベースラインの手法について説明する．ベースラインは Wikipedia 記事中に存在する Infobox から属性値をルールベースにより抽出する．Infobox とは，Wikipedia 記事の主題についての要約情報を提供することを目的とした，記事の右上に配置する形の規定フォーマットの表である．Infobox の例を図 4 に示す．

1 : https://archive.org/download/cirrussearch-20171106/

ルールベースによる手法の概要図を以下の図 6 に示す．具体的にルールベースによる手法は以下の 2 つの手順で構造化をする．

図 4　トヨタ自動車の Infobox

- 属性名の同定

例えば,「本拠地国」属性の属性値を抽出するとき，Infobox 中に「本拠地国」という属性があれば，その属性の値を参照すればよいが，属性がない場合がある．そのため，Infobox 中にある属性値の中で，学習データから得られる属性値リストにから属性名を同定する．図 5 の例では,「本拠地国」属性について属性値を抽出したい場合は,「本社所在地」属性を参照する必要がある．

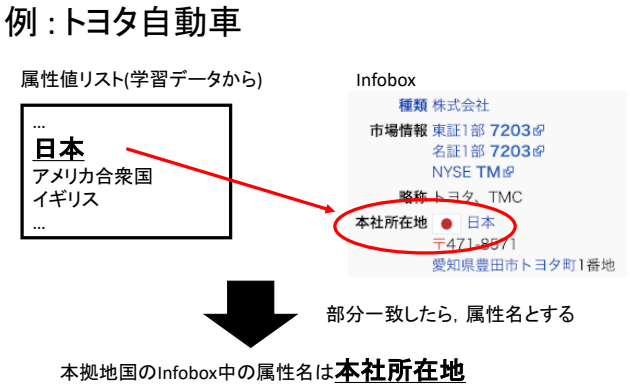

図 5　ルールベースによる属性名の同定方法

- 属性値の抽出

属性名の同定によって得られた属性名に対して属性値を参照して抽出を行う．CirrusSearch のダンプデータでは Infobox 中に国名が国名コードによって記述されている場合がある．図 6 の例では，Infobox 中において「日本」は「JPN」という国名コードで記載されている場合がある．そのため，国名コードに対応した国名のリストを用意し，国名に変換した．

4.3　ベースラインと提案手法を組み合わせた属性値抽出

4.4 節の表 2，3 において，適合率はベースラインによる手法が，再現率，F 値は 3.3 節の記事本文とカテゴリの両方を用いた手法が 1 番高い結果となった．そのため，この 2 つを組み合わせた手法でも実験を行う．組み合わせる方法については 3.3 節と同様に，1 つの手法で属性値が抽出できなければ，別の手法を用いる．

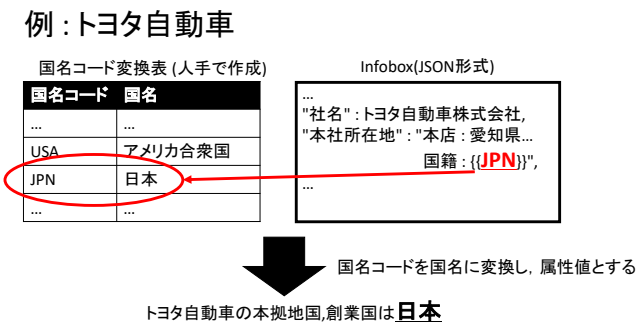

図 6　ルールベースによる属性値の抽出方法

4.4　実 験 結 果

ベースラインによる手法と提案手法を比較した結果を表 2，表 3 に示す．表 2 は「本拠地国」属性，表 3 は「創業国」属性による実験結果である．また，手法の番号については，

- (1)：Wikipedia 記事本文で最初に出現したもの (3.1 節の (1))
- (2)：Wikipedia 記事本文に最も多く出現したもの (3.1 節の (2))
- (3)：Wikipedia カテゴリで最も多く出現したもの (3.2 節)
- (1) → (3)：(1) の手法で属性値がなければ，(3) の手法で抽出 (3.3 節)
- (3) → (1)：(3) の手法で属性値がなければ，(1) の手法で抽出 (3.3 節)
- ベースライン → (3) → (1)：ベースラインの手法で属性値がなければ，(3) → (1) の手法で抽出 (4.3 節)
- (3) → (1) → ベースライン：(3) → (1) の手法でなければ，ベースラインの手法で抽出 (4.3 節)

とする．

表 2 の本拠地国属性での結果において，適合率はベースラインの手法が，再現率と F 値はベースライン→ (3) → (1) の手法が 1 番高い結果となった．表 3 の創業国属性での結果において，適合率はベースラインの手法が，再現率と F 値はベースライン → (3) → (1) の手法が 1 番高い結果となった．

表 2　実験結果 (属性：「本拠地国」)

手法	適合率	再現率	F 値
ベースライン	0.998	0.780	0.875
(1)	0.877	0.747	0.806
(2)	0.850	0.724	0.782
(3)	0.956	0.790	0.865
(1) → (3)	0.878	0.879	0.879
(3) → (1)	0.928	0.930	0.929
ベースライン → (3) → (1)	0.938	0.974	0.955
(3) → (1) → ベースライン	0.931	0.967	0.949

表 3　実験結果 (属性：「創業国」)

手法	適合率	再現率	F 値
ベースライン	0.859	0.787	0.821
(1)	0.745	0.744	0.745
(2)	0.723	0.721	0.722
(3)	0.803	0.779	0.791
(1) → (3)	0.740	0.869	0.799
(3) → (1)	0.785	0.922	0.848
ベースライン → (3) → (1)	0.793	0.965	0.871
(3) → (1) → ベースライン	0.788	0.959	0.865

4.5 考　　察

表 2，表 3 の両方とも，適合率についてはベースラインの手法が 1 番高い．しかし，再現率についてはベースラインを組み合わせたものを除くと，(3) → (1) の手法が 1 番高い．これは，ベースラインの手法が Infobox から抽出しているが，Infobox が存在しない記事については属性値を抽出できないことによるものであると考えられる．また，ベースラインの手法では，あらかじめ国コードに対応した国名のリストを用意する必要があり，このリストは学習データからは作成できない点も注意する必要がある．

表 2，表 3 の両方とも，記事本文中からの抽出方法については (1) の手法が (2) よりも適合率，再現率，F 値の全てが高い．これは日本語版 Wikipedia の記事において，海外の企業についての記事では日本ではどのような展開を行なっているのかが記述されているため,「日本」が最も多く出現していると考えられる．図 2 のスターバックスを例に挙げると，正解となる本拠地国, 創業国属性の属性値は「アメリカ合衆国」だが，本文中では「日本」が最も多く出現した．

(1) と (3) の手法を比較した場合は，(3) の手法が (1) よりも適合率，再現率，F 値の全てが高い．これは 3.2 節に示した，「日本語版の Wikipedia の記事における企業名の記事では，その企業が日本の企業であれば，国について本文中では「日本」と述べず，都道府県等で場所を示す場合が多い」という仮定が正しいものであったと考えられる．これにより，3.3 節の記事本文とカテゴリの両方を組み合わせた手法については，まず (3) の手法で抽出したのちに，(1) の手法を行う (3) → (1) が有効である．

表 2 において，ベースラインと比較して再現率が 1 番高い (3) → (1) の手法において，再現率が向上した要因である (3) → (1) の手法でのみ正解となった属性値の一部を表 4 に示す．ベースラインでは属性値が抽出できない場合が多く，これは記事中に Infobox がないことや，Infobox 中に本拠地国に関する記述がなかったことを示しており，カテゴリと本文による抽出が再現率の向上に有効であったと考えられる．

表 5 は (3) → (1) の手法では正解とならなかった抽出結果の一部を示す．(3) → (1) の手法では「アメリカ」が抽出されている場合において，Wikipedia カテゴリでは「アメリカ合衆国」ではなく「アメリカ」と記述されているため，ベースラインよりも適合率が低下したと考えられる．また，正解は属性値を抽出していないが，(3) → (1) の手法では抽出している場合がある．これは正解となる学習データが Wikipedia の本文と Infobox を見て人手でアノテーションしているため，カテゴリを考慮していないことにより，ベースラインよりも適合率が低下したと考えられる．

ベースラインと提案手法を組み合わせた結果については，表 2，表 3 より，本拠地国と創業国のどちらともベースライン → (3) → (1) の手法の方が (3) → (1) →ベースラインの手法よりも適合率，再現率，F 値の全てが高い結果となった．複数の手法を組み合わせる場合においては，先に用いる手法で抽出できなかった場合のみ，次の手法を用いるため，最初に用いる手法の適合率を重視する必要がある．そのため，適合率が高い方法から抽出を行うベースライン → (3) → (1) の手法が高い結果となったと考えられる．

表 4　(3) → (1) の手法が正解となった抽出結果の一部 (属性：「本拠地国」)

記事タイトル	正解	ベースライン	(3) → (1) の手法
朝日出版社	日本	なし	日本
西日本電線	日本	なし	日本
Rayark	台湾	なし	台湾
ドイチェヴェルケ	ドイツ	なし	ドイツ

表 5　(3) → (1) の手法が正解とならなかった抽出結果の一部 (属性：「本拠地国」)

記事タイトル	正解	ベースライン	(3) → (1) の手法
IBM	アメリカ合衆国	アメリカ合衆国	アメリカ
マーベル・コミック	アメリカ合衆国	アメリカ合衆国	アメリカ
ヤマテ工業	なし	なし	日本
宇宙技術開発	なし	なし	日本

5　まとめと今後の展望

本稿では，森羅プロジェクトにおいて拡張固有表現カテゴリ「企業名」に分類された Wikipedia 記事 600 件に対して，Wikipedia カテゴリと記事本文から国に関する属性の構造化手法を提案した．実験の結果，ベースラインで回答が得られなかったものに対して提案手法を用いることで，ベースラインと比較して再現率，F 値が向上したことを報告した．

今後の展望としては，Wikipedia 記事を 1 つの文書として扱い，どの国の記事なのかを分類する文書分類の問題として考え，SVM やランダムフォレストなどの文書分類手法を用いて構造化を行う．また，本稿では実験の対象とする属性を「本拠地国」と「創業国」に限定していたが，他の属性についても，同様の実験ができないかと考えている．

文　　献

[1] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, Vol. 6, No. 2, pp. 167–195, 2015.

[2] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. Yago3: A knowledge base from multilingual wikipedias. In *CIDR*. www.cidrdb.org, 2015.

[3] G. Miller. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] 関根聡, 小林暁雄, 安藤まや, 乾健太郎. 拡張固有表現に基づく wikipedia 項目の分類と構造化.

[5] 関根聡, 安藤まや, 松田耕史, 鈴木正敏, 乾健太郎. 「拡張固有表表現+wikipedia」データ. 言語処理学会 第 22 回年次大会 発表論文集, 2016.

[6] 関根聡, 小林暁雄, 安藤まや, 馬場雪乃, 乾健太郎. Wikipedia 構造化データ「森羅」構築に向けて. 言語処理学会 第 24 回年次大会 発表論文集, 2018.

[7] Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. Neural joint learning for classifying wikipedia articles into fine-grained named entity types. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Posters*, pp. 535–544, 2016.