複数エンティティがまとめて記述される可能性を考慮したWikipedia記事のクラス分類

鈴木希望 [1]　吉岡真治 [2]
[1] 北海道大学大学院情報科学院　[2] 北海道大学大学院情報科学研究院
zbnadrsplsgame@eis.hokudai.ac.jp　yoshioka@ist.hokudai.ac.jp

概要

Wikipediaの情報に基づいて、構造化した知識を抽出する研究は多く行われているが、基本的には、Wikipediaの記事が一つのエンティティについて述べていることを前提としている。しかし、実際のWikipediaの記事には、小説を原作としたような映画のように、元の創作物からの派生的なエンティティがまとめて記載されることが多い。本研究では、そのような複数エンティティが記述されている可能性を考慮した中で、Wikipedia記事のクラス分類を行う方法を提案するとともに、森羅のデータを用いて本手法の性質について議論する。

1　はじめに

Wikipediaとは世界最大級のインターネット百科事典であり、記事のメタデータとして、概要文、Wikipediaカテゴリ、インフォボックス（テンプレートにより生成）を有する。しかし、Wikipedia自体には、これらの情報を構造化し利用可能にする枠組は存在しないため、これらの知識を構造化するための手法が提案されている。これらの手法には、Wikipediaと同様の編集者による編集作業により構造化を行うWikidata[1)]や、Wikipediaに記述されている知識の自動抽出を試みるDBpedia[1]、BabelNet[2]、YAGO4.5 [3]、知識を拡張固有表現というオントロジに基づいて構築する森羅プロジェクト[4]のプロジェクトなどが存在する。これらの自動抽出を行う研究では、基本的にWikipediaの1記事が一つのエンティティや概念に対応すると仮定してクラス分類や属性の抽出といった研究を行っている。

しかし、Wikipediaの記事には、「永遠の0」のような小説や漫画等の創作物の記事に代表されるように、同名の映画等の、元の創作物から派生したエンティティがまとめて記述される場合があり、Wikipediaの1記事が複数のエンティティと対応すると考えるのが不適切な場合が存在する。ただ、このような記事では、概要文で複数エンティティの存在が示唆されることが多いことが確認されている。また、多くの場合、それぞれのエンティティに対応するインフォボックスが存在している。

そこで、本研究では、記事の概要文を用いた複数エンティティに関する情報の発見をするとともに、インフォボックスの存在によって、十分な情報があるかを検証するという枠組みを用いることによって、複数エンティティについて記載しているページとして認識するとともに、それらのクラスを推定することができる手法を提案し、森羅のデータを用いて本手法の性質について議論する。

2　関連研究

2.1　DBpedia

DBpedia[1]とはWikipediaのメタデータから構造化された情報を抽出し、構造化された知識を構築するコミュニティプロジェクトである。

DBpediaでは、Wikipediaの1つの記事が1つのエンティティと考えて属性を付与するために、例えば、地理情報などにおいて、ページ中に記載されている様々な地点の座標情報を対象記事に関する情報として設定するという問題を起こすと言った問題が指摘されている[5]。

2.2　BabelNet

BabelNet[2]とはWikipediaとWordNetとの対応付けを行うとともに、代表的なカテゴリをPageRankによる順位付けによって選択して、クラス分類を行なっているオントロジである。

エンティティの単位としては、DBpediaと同様に

1）https://www.wikidata.org/

Wikipedia の記事の単位が 1 つの単位となっている。一方で 1 つのエンティティに対し、クラスラベルは代表的なカテゴリによって付けられており、代表的なカテゴリが複数ある場合には複数のラベルをつけることが可能である。

2.3 森羅プロジェクト

森羅プロジェクト [4] とは Wikipdia の知識を拡張固有表現に基づき、計算機が可読可能な構造に変換された知識の構築を目的とする 2017 年にスタートしたリソース構築プロジェクトである。このプロジェクトでは Wikipedia の知識をトップダウン手法で知識を構築するにあたって、拡張固有表現及び拡張固有表現階層、と呼ばれる関根の 150 種類の拡張固有表現階層定義 [6] を経て作成された固有表現オントロジーを使用している。この固有表現オントロジを用いて、記事の代表的な拡張固有表現階層の属性情報を抽出し、該当する Wikipedia 記事にその属性値を対応させることによって、構造化された Wikipedia 知識の構築を目標としている。

エンティティの単位としては、DBpedia と同様に Wikipedia の記事の単位が 1 つの単位となっている。一方で 1 つの記事内に複数のクラスに対応する情報が十分に記載されていると判断される場合には、複数のクラスラベルを付与することが許されている。しかし、森羅で利用されているクラス分類については、抽象度が高く排他的な分類である[2]と考えられるため、このような複数クラスに属する記事は、異なるクラスの複数エンティティについて記載している可能性が高いと考えている。

2.4 Wikidata

Wikidata は、人間とコンピュータの双方が平等に参照・編集できる知識データベースサイトである。Wikidata は Wikipedia を含む Wikimedia の姉妹プロジェクトであり、Wikipedia の多言語サービスの記事のハブとしての役割を果たしている。

Wikidata の１エントリはもともと Wikipedia の記事と対応つく形で作られていたが、Wikipedia の記事の単位の問題もあり、Wikipedia とは直接対応づかないエントリも存在する。

１つのエンティティに対し、クラスラベルは人手で付けられ、複数付けることが可能である。

2） Wikipedia カテゴリで用いられる「作家」のような職業を表すクラスは存在せず、人名・組織名といったレベルの抽象度であるため、排他的であると考える。

2.5 YAGO4.5

YAGO は DBpedia と同様に Wikipedia のメタデータを抽出して使用することで、知識を構造化するオントロジであった。YAGO4 以前では、Wikipedia カテゴリをクラス分類に用いていたが、前述したように、カテゴリには、必ずしも記事の分類を表すものではないカテゴリなどが存在することから、YAGO4[7] 以降では、カテゴリを利用していない。エンティティの単位としては、Wikipedia の記事に対応した Wikidata の記事の単位が 1 つの単位が原則となっているが、先に述べた Wikipedia に対応しないような Wikidata に対応づけることも可能である。また、一つのエンティティに対し、クラスラベルは複数つけることが可能である。

3 提案手法

基本方針は、Wikipedia 冒頭の概要文を考慮した複数クラス候補の抽出と、十分なエンティティの情報があるかを検証することで、複数エンティティに関する記述をしている記事の発見と分類を行う。

例えば、「ちょびっツ」[3]の記事においては、記事の概要文の冒頭に、「『ちょびっツ』は、CLAMP による日本の漫画およびアニメ作品」という記述があり、本記事には、「漫画」と「アニメ作品」の二つのエンティティに関する記載があることが示唆される。また、この記事には、漫画とアニメ作品を記述される際に用いられるインフォボックスがそれぞれ用いられているだけでなく、カテゴリにも、「2000 年の漫画」、「SF アニメ」といったそれぞれに対応する情報が存在する。このような記事について、「漫画」:「アニメ」の二つのクラスラベルを付与する。

インフォボックスが存在しない場合には、「花畑運河」[4]における「運河」のように十分な記述が存在しない場合が存在するため、複数エンティティを認定するためには、対応するインフォボックスの存在を必須とすることとした。

本システムでは、JSAI2024 で行なった記事分類 [8] を元に、森羅の ENE カテゴリ、以降はクラスと呼ぶ、の情報をクラス分類の正解として利用することで、次のような分類規則を作成し、記事のクラス分類を行う。なお、元の記事分類手法と重複する分類規則 2.-4. については省略する。

3） https://ja.wikipedia.org/wiki/ちょびっツ
4） https://ja.wikipedia.org/wiki/花畑運河

表 1　各手法におけるエンティティの単位とクラスラベル

	エンティティの単位	クラス
DBpedia	Wikipedia の記事 (各言語)	単数
BabelNet	Wikipedia の記事 (各言語)	複数可
森羅プロジェクト	Wikipedia の記事 (各言語)	複数可
Wikidata	Wikipedia の記事+人手	複数可
YAGO4.5	Wikipedia または Wikidata	複数可
提案手法	Wikipedia の記事 (複数エンティティ存在)	複数可

1. 定義語の抽出
2. 定義語のクラス分類
3. 定義語を用いた代表的なカテゴリの選択と代表語の抽出
4. カテゴリのクラス分類
5. インフォボックスのクラス分類
   インフォボックスについても定義語と同様に森羅のクラスの情報と比較することで、インフォボックスのクラス分類への有用性を検討する。
6. 記事のクラス分類
   インフォボックス、代表語、定義語、カテゴリのうち、クラス分類に有用なものを用いて、記事のクラス分類を行う。

3.1　定義語の抽出

元の記事分類手法と同様にして定義語の抽出を行うが、本システムでは並列を考えた複数を定義語として採用する。具体的には「A・B・C 並びに D」のような並列を表す定義語が取得できた場合は、そこから「A」、「B」、「C」、「D」のように「ならびに」のような並列を表す語で分割を行い、元の定義語と置き換える形で採用する。

3.2　インフォボックスのクラス分類

インフォボックスについてクラス分類に有用かどうかを判定する。各インフォボックス $infobox$ に対して、定義語を含む記事が森羅の特定のクラスに属しているかどうかを示す指標である $precision(infobox, eneid)$ を計算する。$num(infobox) \geqq 5$、$precision(infobox, eneid) \geqq 0.9$ となるようなクラスが存在するテンプレートを、有用なテンプレートとする。

$$precision(infobox, \text{クラス}) = \frac{|P_{infobox} \cap P_{eneid}|}{|P_{infobox}|} \tag{1}$$

P_{all}: 記事の全体集合
$P_{infobox}(\subseteq P_{all})$: インフォボックス $infobox$ を持つ記事の集合
$P_{eneid}(\subseteq P_{all})$: 森羅のクラスに属する記事の集合
$num(S)$: 集合 S の要素数を返す関数

3.3　記事のクラス分類

記事に付く分類済みのインフォボックス・定義語・代表語・カテゴリを使い、記事のクラス分類を行う。

具体的には以下の 8 つのパターンで、記事のクラス分類を行う。下位のパターンは上位のパターンでの分類ができなかった場合に適用する。なお、本研究の分類パターン 4-7 は、順に元の記事分類手法の分類パターン 1-4 と同等である。

記事のクラス分類

分類パターン 1　有用なテンプレートが存在し、有用な代表語、定義語およびカテゴリでのクラス分類の結果に、テンプレートでのクラス分類結果との共通する結果があれば、その共通部分をつける

分類パターン 2　有用なテンプレートが存在し、有用な代表語、定義語およびカテゴリでのクラス分類の結果に、テンプレートでのクラス分類結果との共通する結果がなく、代表語による分類が可能である場合には、代表語による分類を行う。

分類パターン 3　有用なテンプレートが存在し、有用な代表語、定義語およびカテゴリでのクラス分類の結果に、テンプレートでのクラス分類結果との共通する結果がなく、代表語による分類が可能でない場合には、テンプレートによる分類を行う。

分類パターン 4　元の記事分類手法の分類パターン 1 と同等。

分類パターン 5　元の記事分類手法の分類パターン 2 と同等。

分類パターン 6　元の記事分類手法の分類パターン 3 と同等。

表 2　森羅のマルチエンティティとシングルエンティティ

	森羅マルチ			森羅シングル			
	完全一致	部分一致	不一致	完全一致	部分一致	不一致	計
提案手法マルチ	450	571	55	0	504	241	1821
提案手法シングル	0	7427	9745	72978	0	172416	

分類パターン 7　元の記事分類手法の分類パターン 4 と同等。

分類パターン 8　上記のパターンに全て当てはまらない場合には、訓練データの中で、上記のパターンに当てはまらなかった記事群を抽出し、その記事群で最も出現頻度の高いクラスをを答とする

4　森羅データを用いた実験

本手法の複数エンティティの分類に関する性能を議論するために、人手で作成された森羅 2023 で配布されたクラス分類タスクの教師データについて、複数のクラスラベルを持つ場合の分類性能に注目して分析を行う。対象となる記事数は、訓練記事数 656298 件で、264,146 件をテストデータとして行った。

結果については、複数エンティティを認識したか否か（提案手法マルチ・提案手法シングル）と森羅の分類が複数のクラスに分類されているか否か（森羅マルチ・森羅シングル）の二つに分類して、結果を示す。

我々の手法で複数クラスと判断されたのは、1821 件のみで、森羅の 18,248 件に比較して少ない状況である。これは、我々の複数エンティティに関する認識に関する条件が厳しいため、多くの複数エンティティを見つけ損なっていることが確認できる。これらの記事では、インフォボックスに対応するテンプレートの存在は確認していることが多いものの、現在の定義語の抽出の際に、簡単な分割しか行っていないため、「A, およびそれを原作とした B」などの並列構造を認識できていないために、複数エンティティの発見が失敗したのではないかと考えている。

また、見つけた場合においても、部分一致が多い。一方、「ジュ・テーム・モワ・ノン・プリュ」[5] のように、定義文、カテゴリ、インフォボックスが「音楽名」「映画名」に対応する形で存在するために、二つのラベルをつけたが、森羅には、「音楽名」しか存在しないといった事例も存在した。森羅の基準は、判定者の主観によるところがあるため、我々の客観的な分析とは異なるデータが存在することも確認できた。

また、表に森羅と本手法のそれぞれにおいて、複数クラスがついた記事のうち、その組み合わせの多いもの上位 5 件を記す。

Wikipedia には、両方で上位にある「書物名」と「番組名」や「チャンネル名」と「企業名」のような派生的なデータを持つようなページが一定数存在することが確認できる。一方で、「城名」と「遺跡名__その他」のような同じエンティティが時間的変遷により、クラスが変わるような事例も存在し、その取り扱いについても、今後検討していく必要がある。

表 3　森羅の複数クラスの組み合わせ上位 5 件

森羅マルチ	記事数
書物名, 番組名	2036
城名, 遺跡名__その他	1524
映画名, 書物名	1188
停車場名, 商業施設名	1054
チャンネル名, 企業名	961

表 4　本手法の複数クラスの組み合わせ上位 5 件

本手法	記事数
映画名, 番組名	327
書物名, 番組名	291
チャンネル名, 企業名	272
映画名, 書物名	208
植物名, 食べ物名__その他	57

5　おわりに

本研究は、複数の排他的なクラスと判断されるクラスを複数有する記事を複数エンティティについて記載している記事として認識するとともに、それらのクラスを推定することができる手法を提案した。今後は、より柔軟に定義語の候補を抽出する手法の検討を進めるとともに、本手法の洗練化を図っていきたい。

5）https://ja.wikipedia.org/wiki/ジュ・テーム・モワ・ノン・プリュ

参考文献

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point for the web of data. **Journal of Web Semantics**, Vol. 7, No. 3, pp. 154–165, 2009. The Web of Data.

[2] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. **Artificial Intelligence**, Vol. 193, pp. 217–250, 2012.

[3] Fabian M. Suchanek, Mehwish Alam, Thomas Bonald, Lihu Chen, Pierre-Henri Paris, and Jules Soria. Yago 4.5: A large and clean knowledge base with a rich taxonomy. In **Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '24, p. 131–140, New York, NY, USA, 2024. Association for Computing Machinery.

[4] 関根聡, 中山功太, 野本昌子, 安藤まや, 隅田飛鳥, 松田耕史. 拡張固有表現に分類された 31 言語の wikipedia 知識ベース. 言語処理学会 第 28 回年次大会 発表論文集, 2022.

[5] 吉岡真治. Wikipedia を中心とした linked open data に関する一考察. 研究報告情報基礎とアクセス技術（IFAT）, Vol. 2012, No. 1, pp. 1–5, 07 2012.

[6] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In Manuel González Rodríguez and Carmen Paz Suarez Araujo, editors, **Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)**, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association (ELRA).

[7] Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. Yago 4: A reason-able knowledge base. In Andreas Harth, Sabrina Kirrane, Axel-Cyrille Ngonga Ngomo, Heiko Paulheim, Anisa Rula, Anna Lisa Gentile, Peter Haase, and Michael Cochez, editors, **The Semantic Web**, pp. 583–596, Cham, 2020. Springer International Publishing.

[8] 鈴木希望, 吉岡真治. Wikipedia カテゴリと定義文を利用した記事のクラス分類. 人工知能学会全国大会論文集, Vol. JSAI2024, pp. 4Xin284–4Xin284, 2024.