「拡張固有表表現＋Wikipedia」データ

(2015 年 11 月版 Wikipedia　分類作業完成版)

関根聡 [1) 2)]　安藤まや [1)]　小林暁雄 [2)]　松田耕史 [3)]　鈴木正敏 [3)]　Duc Nguyen[4)]　乾健太郎 [2) 3)]

1) ランゲージ・クラフト　2) 理研 AIP　3) 東北大学　4) オルツ

{sekine, ando}@languagecraft.com, {matsuda, m.suzuki, inui}@ecei.tohoku.ac.jp

{satoshi.sekine, akio.kobayashi}@riken.jp, nguyen.tuan.duc@alt.ai

1. はじめに

Wikipedia の約 73 万項目を 200 種類の拡張固有表現に分類したデータが完成した。

これは、約 2 万項目のトレーニングデータを元に機械学習を利用して自動分類した結果を人手で確認、訂正したデータであり、その精度は極めて高いものである。Wikipedia の原データは 2015 年 11 月版であり、約 95 万ある項目の内、比較的重要ではない項目の分類作業を省くために、被リンク数が 5 以下の項目は分類対象とはしていない。

本データ作成のより詳しい背景や意義および約 2 万項目のトレーニングデータの分析結果については[関根ら 16]に、機械学習については[鈴木ら 16][Suzuki et al. 18]に報告している。本論文では、主に完成版の詳細について報告する。

2. 背景と目的

自然言語理解を実現するためには、言語的及び意味的な知識が必要なことは論を待たない。しかしながら、大規模な知識の作成は非常に膨大なコストがかかり、メンテナンスも非常に難しい問題である。名前を中心とした知識において、 クラウドによって作成されている Wikipedia はコストの面でもメンテナンスの面でもそれ以前の知識ベースと呼ばれるものの概念を一新した。しかし、この Wikipedia を自然言語処理のための知識として活用しようと考えると障壁は高い。Wikipedia は人が読んで理解できるように書かれており、計算機が利用できるような形ではないためである。計算機の利用を念頭において構築された知識には、CYC、DBpedia、YAGO、Freebase、Wikidata などがあるが、それぞれに解決すべき課題があると考えている。特に CYC ではカバレージの問題、他の知識では、首尾一貫した知識体系に基づいていない構造化の問題が重要であると考える。より詳しい関連研究の分析は[関根ら 17]に詳しい。これらの課題を解決するため、私たちは、名前のオントロジー「拡張固有表現」[Sekine 08]に Wikipedia 項目を分類し、拡張固有表現に定義されている属性情報を抽出することで計算機利用可能な Wikipedia の構造化データ「森羅」の構築を進めている(図 1)。

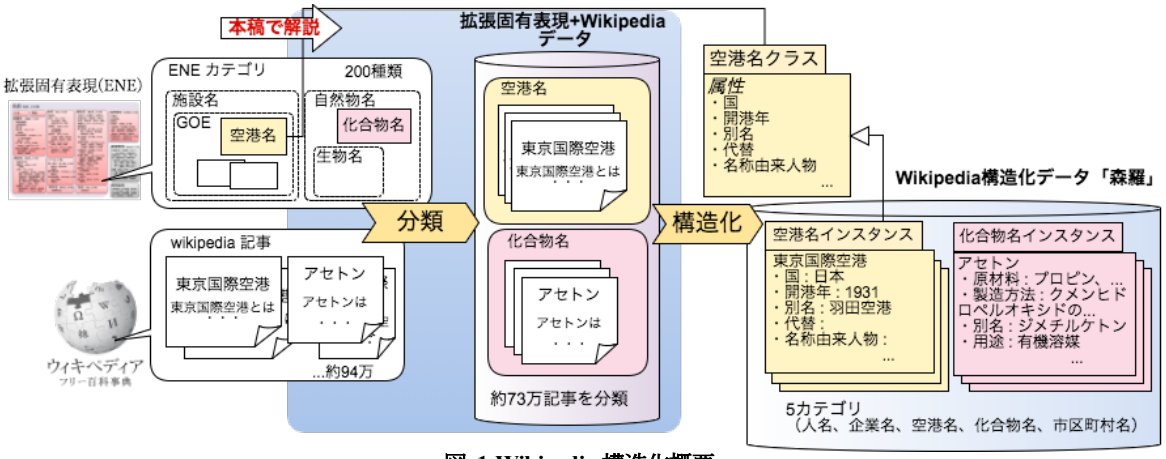

図 1:Wikipedia 構造化概要

本稿では、構造化のための第一歩である、Wikipediaの項目の意味的な分類作業とその結果について解説する。構造化については[関根ら 18]にて解説する。拡張固有表現は、百科事典や既存の質問応答システム、概念辞書を利用して作成されており、200種類のノードが定義さている。このため、Wikipediaに収録された、幅広い名前のエンティティを高カバレージで分類することが可能である。上記に挙げたCYC以外の知識ベースでは、カテゴリの設定がクラウドによってバラバラになされており、エンティティの分類体系として使用することは困難である。拡張固有表現の定義に従った拡張固有表現抽出システムの試み[新納ら 06][Duc 17]や拡張固有表現タグ付きのテキストデータも公開されている[橋本ら 08]。しかしながら、概念の種類が非常に多いため、抽出システムは十分な精度が出ていない。この一つの原因としては、スパースなトレーニングデータを補完することのできる拡張固有表現辞書がないことが挙げられる、本論文で述べるデータはこの目的のためにも有効である。

3. 拡張固有表現

拡張固有表現とは、[Sekine 08]によって定義された固有表現に関する定義であり階層構造を持つ。人名、地名、組織名だけではなく、イベント名、役職名、芸術作品名などの新しい固有表現や、地名に含まれる河川名などの地形名や星座名などの天体名などが含まれる。Version 7.1.0では最大3階層までの全部で200種類の拡張固有表現が定義されている。 [ENE definition HP]

4. データの説明

本章では、作成したデータの説明を行う。まず、対象とした2015年11月版のWikipediaの紹介をし、その中で分類対象となったデータの位置付けを明確にする。その後、サンプルデータを示しデータ内容の概要を説明する。最後に、分類結果の全体的な統計データを示す。

4.1. Wikipedia項目の統計データ

本データの特徴を明らかにする為に、Wikipedia項目(2015年11月版)と本データに関する統計データを表１に示す。 表にある通り、Wikipediaの全項目数は1,588,284項目であるが、項目が実際のエンティティではないメタな項目（転送、リスト、曖昧性回避項目など）が約64万項目あり、実際のエンティティが書かれているとされる項目は、943,508項目である。そして、その中で、メジャーな項目を判断する為に今回採用した被リンク数が5以上の項目は782,517項目あり、本データはこの約78万項目を拡張固有表現に分類したデータである。

項目タイプ	被リンク数	データ数
Normal	５以上	782,517
	５未満	160,991
	Normal 合計	*943,508*
Redirect		594,035
List		10,364
Disambiguation		40,377
メタ項目合計		*644,776*
全項目の合計		*1,588,284*

表１．Wikipedia項目の統計データ

4.2. データの具体例

データの具体例を図２に示す。ここでは「ルイ・ヴィトン」の項目に対して、Wikipediaに存在する様々な情報をJSON形式でまとめ、そして、拡張固有表現の分類がENEsの項に示されている。この例では、201604の自動分類では「企業名」「製品名__その他」、201703の自動分類では「製品名__その他」のみに分類され、最終的には201712の人手の分類によって「人名」と「製品名__その他」に分類されている（ブランド名は「製品名__その他」であり、LVMHが企業名となる）。ここで分類作業のタグは、以下のような意味である。最初の部分は、“AUTO”（自動分類）か“HAND”（人手作業)のいずれかであり、AUTOの場合には、TOHOKU_201604（初期の2万項目を基にした自動分類）かTOHOKU_201703（2017年3月当時のデータを基にした自動分類）である。HANDは、LC_2015, ALT_201703, LC_201712のいずれかである。

図２.データの具体例（「ルイ・ヴィトン」）

{"SID":161377,
"wikipedia_ID":"260264",
"entry":"ルイ・ヴィトン",
"clean_entry":"ルイ・ヴィトン",
"page_property":"Normal",
"redirect_to":"",
"redirect_from":["ヴィトン","ルイヴィトン","Vuitton","Louis vuitton","LOUIS VUITTON","ルイビトン","ルイ・ビトン"],
"link_from_N":237,
"link_anchor":[{"count":1,"anchor":"ルイ・ヴィトン・ジャパン・カンパニー"}, {"count":1,"anchor":"ヴィトン"}, {"count":7,"anchor":"Louis Vuitton"}, {"count":1,"anchor":"ルイヴィトン"},{"count":224,"anchor":"ルイ・ヴィトン"}, {"count":1,"anchor":"LVMH ルイ・ヴィトン"}, {"count":1,"anchor":"LOUIS VUITTON"}, {"count":1,"anchor":"ヴェトン"}],
"category_info":["フランスのファッションブランド","フランスの服飾関連企業","パリの企業","LVMH"],
"first_sentence":"ルイ・ヴィトン()は、フランスのマルティエ(スーツケース職人)であるルイ・ヴィトン(、1821年8月4日 - 1892年2月28日)が創始したファッションブランド。",
"abstract":"ルイ・ヴィトン()は、フランスのマルティエ(スーツケース職人)であるルイ・ヴィトン(、1821年8月4日 - 1892年2月28日)が創始したファッションブランド。LVMH(モエ・ヘネシー・ルイ・ヴィトン)グループの中核ブランドである。LVMHの2008年の売上高は239億ドル。服飾部門における2014年現在のデザイナーはニコラ・ジェスキエール。またメンズコレクションについてはキム・ジョーンズがチーフディレクター。",
"listed_in":["フランスの企業一覧","慶應義塾大学の人物一覧","ファッションブランド一覧","フランス人の一覧","灘中学校・高等学校の人物一覧"],
"ENEs":[{"ENE":[["企業名",0.6093719005584717],["製品名_その他",0.9585748314857483]],"annotation_flag":"AUTO.TOHOKU_201604"},{"ENE":[["製品名_その他",0.8671039342880249]],
"annotation_flag":"AUTO.TOHOKU_201703"}, {"ENE":[["人名","1.0"], ["製品名_その他","1.0"]], "annotation_flag":"HAND.LC_201712"}]}

4.3. 作成データの統計

本節では、今回作成したデータの 782,517 項目の統計情報を示す。 表２に、タグの付与を行った項目の内、拡張固有表現ではなく、主に一般名詞などの概念を表している CONCEPT と、転送、曖昧性回避、リストページなど Wikipedia メタ情報の IGNORED のタグがつけられたデータ数を示す。

表２．作成データの統計情報

説明	データ数
分類対象全データ数	782,517
IGNORED	11,024
CONCEPT	43,142
拡張固有表現のデータ	728,255

表１で説明したメタ情報は、Wikipedia 自身の記法に基づき Redirect などが明示的に示されているものである。しかしながら、正式な記法にのっとらず、普通の項目のような形で書かれているが実はリストや曖昧性回避であるといったページも数多く存在するため、ここでもそのような項目が分類されている。

次に、表３に拡張固有表現に分類された項目数の多いものを挙げる。複数の拡張固有表現に分類された項目は全体の約 2%の 13,967 項目存在する。

表３．高頻度の拡張固有表現

人名	227,228	文学名	17,854
市区町村名	41,735	映画名	16,537
音楽名	37,675	電車駅名	16,154
製品名_その他	31,873	道路名	15,218
番組名	30,276	競技会名	14,504
企業名	25,442	主義方式名_その他	13,789
学校名	22,325	公演組織名	9,592

逆に、低頻度の拡張固有表現の頻度に対する種類数と具体例を表５に示す。頻度の低いものは主に数値表現、時間表現、アドレス等であり、これらが Wikipedia の項目に挙がらないことは理解できる。また、被リンク数で分類の閾値を設定したため、有名な項目が少ない種類の拡張固有表現は頻度が小さくなっている。実際に拡張固有表現の全種類の 200 の内、約半数が頻度 700 以下であることが分かった。今後の機械学習による自動分類を行う際には、このような頻度の低い拡張固有表現に関するトレーニングデータの作成方法についてしっかり検討する必要がある。

表５．低頻度の拡張固有表現

頻度	拡張固有表現数	例
0	25	重量、カロリー、電子メイル
1	8	緯度経度、郵便住所、週数期間
2-5	13	年齢、電話番号、期間_その他
6-10	3	曜日表現、学齢、URL
11-20	3	年数期間、時刻表現、自然色名

5. 拡張固有表現分類チェック時の注意点

分類チェック時に発見された注意点を紹介する。

判断の材料

文中に出現する固有表現に対してその種類を同定する問題は文脈情報を使うことによって解決可能である場合が多い。しかし、文脈を持たずに項目そのものを与えられただけではこの曖昧性は解決されない。我々は、基本方針として、分類されるべきカテゴリは、可能性のある拡張固有表現を網羅的に挙げるのではなく、その項目が一般的に想起される範囲に限ることとした。この範囲を客観的に決めることは難しく、Wikipedia のページをすべて読むことは時間がかかる。試行錯誤の結果、この問題は Wikipedia の構造に頼ることとし、内容を確認する箇所は、原則としてインフォボックス、冒頭部分、目次の 3 か所に限定した。

チェック作業の高精度化、効率化

78 万項目のチェックには時間がかかる。そのため、以下の 2 つの方法で効率化を図った。ひとつは、「定義語」というものを導入し、定義語ごとのチェック作業の実施である。定義語とは、Wikipedia の 1 文目で見出し語を定義していることばを指す。例えば見出し語「安倍晋三」の 1 文目には「安倍晋三（あべしんぞう、1954 年（昭和 29 年）9 月 21 日-）は、日本の政治家。」と書かれており、【政治家】という表現で安倍晋三は定義されている。これが定義語となる。結果的に、定義語は人手による拡張固有表現チェック作業を効率化するために最も有効な情報となった。それは、同じ定義語で説明される見出し語は同じ拡張固有表現に分類される可能性が高いためである。具体的には、例のように「政治家」と定義されている見出し語はたいてい人名となる。このような定義語を自動抽出し、定義語ごとにファイルを分割し、チェック作業を行った。これにより作業の効率化だけではなく、分類の揺れの減少に役立ったと考えている。

さらに、同じ定義語を持つデータセットの内、定義語の頻度が高く、機械学習によって付与された拡張固有表現の揺れが少なく、尚且つ機械学習が計算した信頼度が 0.99 より高いデータ（最高値は 1）は、以下の手法を用いてチェック作業を簡略化した。そのデータセットの中からランダムに抽出した 200 件を人手でチェックし、異なった拡張固有表現のカテゴリとなるものが 1%未満（つまり 0 件か 1 件）であれば、機械による拡張固有表現をそのままデータセット内の全ての項目に無条件で採用することとした。作業を簡略化した定義語の例を表 6 に示す。同じ定義語、拡張固有表現のペアでも信頼度が 0.99 未満であれば人手でチェックした。このようにして、チェックを省略できた項目数は約 15 万項目であった。

表５．作業を簡略化した定義語の例

ENE	定義語例	信頼度 0.99 より大きい	データ数
人名	政治家	6,038	6,041
電車駅名	駅	13,291	13,425
道路名	一般県道	6,308	6,380
音楽名	シングル	12,579	12,618
市区町村名	村	9,003	9,017

システマティック・ポリセミー

見出し語には複数の側面を持つものがある。例えば「サンマ」という魚は、魚類であると同時に、食べ物とみなすことも可能である。こういった多義性は、特定の項目に特有のものではなく、これらと同様のカテゴリの項目に共通してみられる多義性である。これは「システマティック・ポリセミー」呼ばれる多義性の一種である。[Peters and Peters 00]。先に説明したように定義語ごとに拡張固有表現のチェックをすることで、このような多義性の見落としが減少したと考えられる。

一般表現の扱いと区別

何を固有表現とみなし、何をそうでない一般的表現とみなすかも難しい問題である。例えば、「くわがた」は一般的には固有名詞ではないが、拡張固有表現に分類される。一方「寄生虫」は一般的な表現として拡張固有表現に分類されず、CONCEPT（一般的な表現を分類するタグ）に分類される。

拡張固有表現の分類判断基準は、「あの○○の名前を教えてください」の「○○」に固有表現階層のクラスの名前を入れた質問文を作成し、その答として満足できるような単語をその固有表現クラスに分類するとしている。このテストにおいて「あの○○の種類の名前を教えてください」という表現の方が「あの○○の名前を教えてください」という質問よりも、その単語を答とする質問においては自然であるという場合には、その単語はそのカテゴリに入れないこととした。（上記２つの質問文の○○に「動物／昆虫」を入れた場合、「寄生虫」は「種類の質問文」の方が自然であり、「クワガタ」は「名前の質問文」の方が自然である）チェック作業においてもこの質問文を使用したが、拡張固有表現の定義とも絡み難しい問題である。

拡張固有表現の定義書ではカバー出来てないもの

拡張固有表現のようなオントロジーは、トップダウンに設計できるものではなく、事例を見ながらボトムアップに近い形で設計していく。したがってこれまでも頻繁に更新されてきており、現行の定義ではカバーしきれない部分が存在する（ただし、全ての中間カテゴリには「＊__その他」というカテゴリが用意され、分類できない項目は存在しない）。例えば、今回の作業により新たに定義を考えうると判断したものには以下のものがある。番組内のコーナーの名前（現状：番組名）、「アップル対アップル訴訟」のような訴訟の扱い（イベント名_その他）、「北京軍区」のような軍区の扱い（国内地域名及び組織名_その他）。

6. Wikipedia の構造化データ構築に向けて

Wikipeda 項目の大規模な分類を完了した。しかし、最終目的は分類ではなく、Wikipedia に書かれている世界知識を計算機が扱えるようにするための構造化である。構造化とはそれぞれのカテゴリごとに定義された属性値を、各項目から抽出しテーブルの形で整理することを意味する。この構築は分類のように人手を全ての項目に介入させる形では実現できず、多くの参加者を巻き込んだプロジェクトの中で実現していこうと考えている[関根ら 18]。

7. まとめ

Wikipedia の約 73 万項目を拡張固有表現に分類した「拡張固有表現＋Wikipedia」データ（2015 年 11 月版 Wikipedia　分類作業完成版）を作成した。このデータは、ランゲージ・クラフトから分類作業の実費負担をいただく形の有償で公開している。（連絡先：info@languagecraft.com）

参考文献

[関根ら 18] 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（2018）

[関根ら 17] 関根聡, 安藤まや, 小林暁雄, 乾健太郎 「拡張固有表現に基づく Wikipedia 項目の分類と構造化」. 人工知能学会　第 43 回セマンティックウェッブとオントロジー研究会(2017)

[関根ら 16] 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(2016)

[鈴木ら 16] 鈴木正敏, 松田耕史, 関根聡, 岡崎直観, 乾健太郎. Wikipedia 記事に対する拡張固有表現ラベルの多重付与. 言語処理学会第 22 回年次大会（2016）

[Sekine 08] Satoshi Sekine. Extended Named Entity Ontology with Attribute Information. LREC08

[新納ら 06] 新納浩幸, 関根聡 拡張固有表現タガーの作成とその問題点の考察, 言語処理学会第 12 回年次大会（2006）.

[橋本ら 08] 橋本泰一, 乾孝司, 村上浩司. 拡張固有表現タグ付きコーパスの構築. 情報処理学会自然言語処理研究会（2008）

[Duc et al. 17] Tuan Duc Nguyen; Khai Mai; Thai-Hoang Pham; Minh Trung Nguyen; Truc-Vien T. Nguyen; Takashi Eguchi; Ryohei Sasano; Satoshi Sekine. Extended Named Entity Recognition API and Its Applications in Language Education

[Higashinaka et al. 12] Ryuichiro Higashinaka, Kugatsu Sadamitsu, Kuniko Saito, Toshiro Makino, Yoshihiro Matsuhiro. Creating an Extended Named Entity Dictionary from Wikipedia. COLING 2012.

[Peters and Peters 00] Wim Peters and Ivonne Peters. Lexicalised systematic polysemy in wordnet. LREC-2000

[ENE definition HP] http://nlp.cs.nyu.edu/ene

[Suzuki et al. 18] Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, Kentaro Inui, A Joint Neural Model for Fine-Grained Named entity classification of Wikipedia Articles, IEICE Transactions on Information and Systems, vol.E101-D, No.1 pp.73-81, 2018.