「拡張固有表表現＋Wikipedia」データ

関根聡[1)]　安藤まや[1)]　松田耕史[2)]　鈴木正敏[2)]　乾健太郎[2)]
1）ランゲージ・クラフト　2）東北大学
{sekine, ando}@languagecraft.com, {matsuda, m.suzuki, inui}@ecei.tohoku.ac.jp

1. はじめに

大規模なテキストデータから、言語的及び意味的な知識を獲得して、言語解析に応用する研究が盛んに行われている [Hearst 92][安藤ら 03][Pantel and Pennacchiotti 06] [SK-Symposium 08] [Mikolov et al. 13]。これらの研究に基づいた技術は、大量なテキストデータが出現してから目覚ましい進歩を遂げ、その成果に基づいて様々な知識が整備されてきた。しかしながら、言語処理の現実的な応用を考えた時には、そのような知識ではカバレージと精度に限界があるため、より大規模で精度の高い知識が求められているのが現状である。上記に挙げた限界には、特に、単語の意味的な曖昧性を考慮しない統計的扱いやテキストの分野や文脈を考慮しない扱いに起因する精度の問題や、頻度の低い単語や表現の問題である。これらの手法によって作成された知識の補完を考えた際に、有効な手段の一つとして、構造化された知識をあらかじめ用意し、それを使って言語処理を行うことが考えられる。この 2 つの手法は、言語習得において、ボトムアップに第一言語を習得する方法と、すでに世界のあり方を把握している人が、その知識をトップダウンに使って第二言語を習得する方法と対比できるかもしれない。

本論文では、トップダウンに作成する知識の対象として「固有表現」を扱う。固有表現は、情報抽出の研究から作成されたものであることからも分かるように、現実的に人が知りたい物事の多くのタイプを包含している。その概念は、情報抽出だけではなく、ファクトイド型質問応答、自動要約、機械翻訳、対話処理などでも重要な技術として認識され、基本的な固有表現抽出のシステムは様々な言語で開発され利用されている。ただし、現状では多くのシステムにおいて対象とする固有表現の種類が、人名、地名、組織名、時間表現、数値表現などに限られており、より幅広く応用するには不十分であるという現実がある。

この問題に鑑み、[Sekine 08]は、新聞記事に書かれた固有表現を幅広くカバーする「拡張固有表現」を提案した。この拡張固有表現は 200 種類の概念を含み、百科事典や既存の質問応答システム、概念辞書に基づいて定義された。こ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]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、抽出システムは高精度の結果を出せていない。この一つの原因としては、スパースな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。

この辞書の問題は著作権も絡み非常に根深いものではあるが、特に固有表現を対象としたクラウドソーシングで作られた Wikipedia は、この問題を解決する一つの有力なリソースであると言える。2015 年 11 月現在、日本語版の Wikipedia には約 150 万の項目が作られ、人々が興味を持つ固有の対象や概念についての記述がされている。特に、項目の大部分は固有表現であり、これらにトップダウンで定義した分類のタグを付与すること、そして、それぞれの種類に応じた知識の属性に基づいた構造化をすることにより、様々な言語処理への応用が期待できる。[1] [2]

本論文では、Wikipedia の項目に拡張固有表現をタグ付けした「拡張固有表現＋Wikipedia」データを紹介する。本データの作成は開始されたばかりで、現在は約 2 万項目の人手でのタグ付けを終え、それに基づいた機械学習による残り項目の自動タグ付けのデータを作成したばかりである。自動タグ付けの方法については[鈴木ら 2016]を参考にされたい。自動タグ付けの精度は F 値が約 88%であり、今後はそ

[1] ただし、拡張固有表現辞書は[Higashinaka et al. 12]などで作成されているが公開されているものはない。
[2] Wikipedia のカテゴリーや DBPedia は、意味的分類の定義を含むが、不特定多数の個人によるボトムアップな定義であり、言語処理の応用に利用できるとは言い難いレベルである。

の結果を人手で修正し、より精度の高い知識にしていく予定である。また、各拡張固有表現に対して属性が定義されており、その属性定義にも修正を加えつつ、それぞれの項目の情報を属性値の形で整理し構造化された知識も作成していきたいと考えている。

2. 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]

3. データの概要

本節では、「拡張固有表現＋Wikipedia データ」の説明を行う。本データは 2 つの JSON フォーマットのデータからなっている。

構造化 Wikipedia データ

Wikipedia の全項目に関して言語処理の応用のために重要な情報を JSON 形式で構造化したデータである。構造化したデータの属性は以下の通りである。

表１．構造化 Wikipedia データの項目

属性名	説明
SID	本データにおける ID
wikipedia_ID	Wikipedia の ID
entry	項目名
clean_entry	標準化された項目名
page_property	ページの種類
redirect_to	リダイレクト先
redirect_from	リダイレクト元
link_from_N	被リンク総数
link_anchor	リンク元のアンカー文字列
category_info	Wikipedia のカテゴリ情報
first_sentence	説明文中の最初の文（自動抽出）
listed_in	リストされている一覧ページ

拡張固有表現タグデータ

Wikipedia 項目の拡張固有表現の情報である。2 つのデータは同じ SID がついていることから関連付けが可能である。

表２．拡張固有表現タグデータ

属性名	説明
SID	本データにおける ID
ENE	拡張固有表現情報
annotation_flag	アノテーション情報

「拡張固有表現(ENE)」は、関根の拡張固有表現 7.1.0 に準拠している。複数のタグがつく場合もあるため、値の形式は配列を使っている。また、CONCEPT（固有表現ではなく、一般的な名詞など）、IGNORED（Wikipedia 固有のメタ情報など、百科事典としての項目ではないもの）のタグが存在する。
「アノテーション情報（annotation_flag)」にはどのような形で振られたタグであるかが記録される。現状では以下の２種類である

- HAND.LC_annotator_201601：ランゲージ・クラフトのアノテーターによる人手でのタグ付け
- AUTO.TOHOKU_201601：東北大学の機械学習システム[鈴木ら 2016]による自動タグ付け

3.1. サンプル

それぞれのデータのサンプルを図１、図２に載せる。ここでは改行を挿入し、見やすい形で表示してある。項目数とデータサイズは、「構造化 Wikipedia データ」が 1,588,284 項目で 110MB、「拡張固有表現タグデータ」が 4.3MB である。


```
{
 "SID": 180417,
 "wikipedia_ID": "259974",
 "entry": "東京都立新宿高等学校",
 "clean_entry": "東京都立新宿高等学校",
 "page_property": "Normal",
 "redirect_to": "",
 "redirect_from": ["新宿高校", "新宿高等学校", "都立新宿高等学校", "東京都立新宿高校", "東京府立第六中学校"],
 "link_from_N": 276,
 "link_anchor": [{"count": 1, "anchor": "都立六中"}, {"count": 1, "anchor": "都立新宿高"}, {"count": 11, "anchor": "東京府立第六中学校"}, {"count": 2, "anchor": "新宿高等学校"}, {"count": 16, "anchor": "新宿"}, {"count": 1, "anchor": "東京都立第六高等学校"}, {"count": 4, "anchor": "東京府立六中"}, {"count": 1, "anchor": "旧制東京都立第六中学校"}, {"count": 2, "anchor": "新宿高"}, {"count": 4, "anchor": "府立六中"}, {"count": 8, "anchor": "都立新宿高校"}, {"count": 8, "anchor": "新宿高校"}, {"count": 216, "anchor": "東京都立新宿高等学校"}, {"count": 1, "anchor": "東京都立新宿高校"}],
 "category_info": ["東京都区部の公立高等学校|しんしゅく", "新宿区の学校|しんしゅくこう", "学校記事"],
 "first_sentence": "東京都立新宿高等学校(とうきょうとりつ しんじゅくこうとうがっこう)は、東京都新宿区内藤町に所在する都立高等学校。",
 "listed_in": ["旧制中等学校・新制高校のナンバースクール一覧", "東京都立新宿高等学校の人物一覧", "東京都高等学校一覧", "旧制中等教育学校の一覧（東京都）"]
}
```


図１．構造化 Wikipedia データのサンプル

```
{
  "SID": 180417,
  "ENE": ["学校名"],
  "annotation_flag":
"HAND.LC_annotator_201511"
}
{
  "SID": 180419,
  "ENE": ["美術博物館名"],
  "annotation_flag":
"AUTO.TOHOKU_201601"
}
{
  "SID": 180420,
  "ENE": ["CONCEPT"],
  "annotation_flag":
"AUTO.TOHOKU_201601"
}
```


図２．拡張固有表現タグデータのサンプル

3.2. 統計

本節では、今回作成したデータの内、被リンク数が 100 以上の 22,677 項目の Wikipedia 項目に対して、人手で拡張固有表現のタグ付けを行ったデータの統計的な情報を報告する。

表３に、タグの付与を行った 22,677 項目の内、拡張固有表現ではない CONCEPT と IGNORED のタグがつけられたデータ数と、拡張固有表現が付与された場合の１つの項目に付与された拡張固有表現の数による分布を示す。

表３．人手で付与されたデータの統計情報

説明		データ数
付与データ数		22,677
IGNORED		621
CONCEPT		2,660
付与された拡張固有表現の数	1	2,1624
	2	850
	3	187
	4	14
	6	2

ほとんどの項目は１つの拡張固有表現が付与されたが、4.6%の項目には複数のラベルが付与された。６つの種類の拡張固有表現が付与された項目は、以下の２つである。

「漢」: 政治的組織名_その他，河川名，郡名，民族名_その他，国名，地域名_その他
「鉄人 28 号」: 番組名，キャラクター名，文学名，製品名_その他，公演名，映画名

次に、拡張固有表現が付与された場合の、高頻度の拡張固有表現を表４に示す。

表４．高頻度の拡張固有表現

人名	4,041	競技会名	625
番組名	2,395	プロ競技組織名	484
企業名	1,701	地位職業名	462
市区町村名	975	映画名	438
製品名_その他	964	公演組織名	363
日付表現	916	学校名	326
文学表現	909	主義方式名_その他	288

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 0 の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、被リンク数の多いものをタグ付与の対象にしたため、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 10 以下であることが分かった。今後の機械学習による自動タグ付与を行う際には、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	55	URL、人数、温度、絵画名
1	8	船名、恒星名、時刻表現
2-5	16	運河名、取引所名、橋名、罪名
6-10	23	地震名、条約名、学齢、湾名
11-20	23	公共機関名、例祭名、国籍名

4. 拡張固有表現タグ付与時の注意点

本節では、「拡張固有表現＋Wikipedia データ」の内、被リンク数が 100 以上の Wikipedia 項目に対して、人手で拡張固有表現のタグ付けを行った作業中に見つかった問題点を紹介する。

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。Wikipedia の説明文にも、いくつのも可能性を提示している場合がある。この問題を突き詰めると例えば、有名な漫画の項目の説明文の最後に「ちなみに、この作品は映画化されたが公開されなかった」とあった場合に映画名を付与すべきかどうかという問題が生じる。我々は、基本方針として、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、

「Wikipedia の 1 文目の記述内容から同定しうる拡張固有表現のみを付与する」ことにした。

システマティック・ポリセミー

例えば、「ワンピース」という作品は、原作は漫画（「文学名」）であるが、TV アニメ化され、映画化され、ゲームにもなっている。個別の出現において、このうちのどの側面に対して注意が向けられているかは、文脈によって異なり、辞書にどれか一つの固有表現ラベルを記載するのみでは不十分である。同様に、「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、あるエンティティ特有のものではなく、これらと同様のカテゴリのエンティティ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]

Wikipedia における偏在

Wikipedia は一般の人が自由に項目を立てられるという特徴から、番組名、ゲーム、ブランド、プログラムング言語などの製品名_その他，体制、方式、規格などの主義方式名_その他に相当する項目が極めて多かった。拡張固有表現の設計の元には小学館の百科事典があるが、それとの差異が感じられる。

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。たとえば、「新幹線」という単語が特定の路線を指すのか、それとも単に高速な鉄道を指すのかは文脈によって異なる。今回は Wikipedia の記事が指すものを分類するという立場から、CONCEPT というタグを作り明示的に一般的な表現を同定したが、拡張固有表現の定義とも絡み難しい問題である。

被リンク数のバイアス

今回は、人手でのタグ付与の元コーパスを選別する基準として、Wikipedia 内での他の記事からの被リンク数という指標を用いた。これは、Wikipedia 全体から記事をランダムにサンプリングすると、多くがマイナーなエンティティになってしまうという問題に対処するための措置である。しかしながら、Wikipedia 内のリンク構造には大きな偏りがあり、例えば、スポーツチームや個別の競技会等の特定のジャンルの記事間は(試合結果等の繰り返し構造を含むため)非常に密にリンクされており被リンク数が大きい。逆に、有名な絵画でも被リンク数は小さく、今回タグ付けしたデータもバイアスを受けている。この問題に対処する方法は明確ではないが、たとえば Wikipedia Page Views のようなより客観的に人々の情報要求を表した統計値を用いることでより客観性の担保を行いやすくなるかもしれない。

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たし、現行の定義ではカバーしきれない部分が存在する。例えば、今回の作業により新たに定義を考えたものには以下のものがある。「黎朝」「南宋」のような王朝の名前、「ミカエル」のような天使の名前、「大阪城」などの城の名前、「バレエ」「ヨガ」のようなダンスや「クラシック」「ジャズ」のような音楽、「将棋」「チェス」のようなゲームの種類の名前、恐竜の名前である。

5. まとめ

Wikipedia の項目に拡張固有表現のタグを付与した「拡張固有表現＋Wikipedia」データを作成した。現在約 2 万項目の人手付与が終わり、機械学習により自動付与された残りの項目の人手チェックを計画している。その後、各項目の属性値データを作成し、構造化された固有表現に関する知識ベースを作成したいと考えている。

参考文献

[Hearst 92] Marti Hearst. Automatic Acquisition of Hyponyms from Large Corpora, COLING92

[安藤 03] 安藤まや、関根聡、石崎俊：定型表現を利用した新聞記事からの下位概念単語の自動抽出，情報処理学会，第 157 回 自然言語処理研究会，pp. 77-82 (2003)

[Pantel and Pennacchiotti 06] Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. COLING/ACL-06

[SK-Symposium 08] NSF Symposium on Semantic Knowledge Discovery, Organization and Use. http://nlp.cs.nyu.edu/sk-symposium/

[Mikolov et al. 13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, Jeff Dean. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems (2013).

[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)

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

[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)

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