Wikipedia 構造化データ「森羅」構築に向けて

関根聡 [1] 小林暁雄 [1] 安藤まや [2] 馬場雪乃 [1)4)] 乾健太郎 [1)3)]
1) 理研 AIP 2) ランゲージ・クラフト 3) 東北大学 4) 京都大学
{satoshi.sekine, akio.kobayashi}@riken.jp , ando@languagecraft.com,
inui@ecei.tohoku.ac.jp, baba@i.kyoto-u.ac.jp

概要

Wikipedia に書かれている世界知識を計算機が扱えるような形に変換することを目的として、Wikipedia を構造化するプロジェクトを推進している。すでに 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ている。このデータをもとに、それぞれの拡張固有表現で定義された属性を個々の項目の説明文やインフォボックスから抽出し、構造化したデータを作成している。

本データの分類部分は[関根ら 18][鈴木ら 16][Suzuki et al. 18]に報告した。本論文では、構造化データ作成についての現状を報告する。

1. 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、 クラウドソーシング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百科事典の概念を一新した。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た知識ベースには、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、他の知識ベースでは、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。こ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 記事を分類し、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 の構造化を進めている(図 1)。本稿では、[関根ら 18]にて報告した分類結果に基づき、Wikipedia ページから属性値を抽出することで構造化データ「森羅」を構築する試みについて解説する。

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

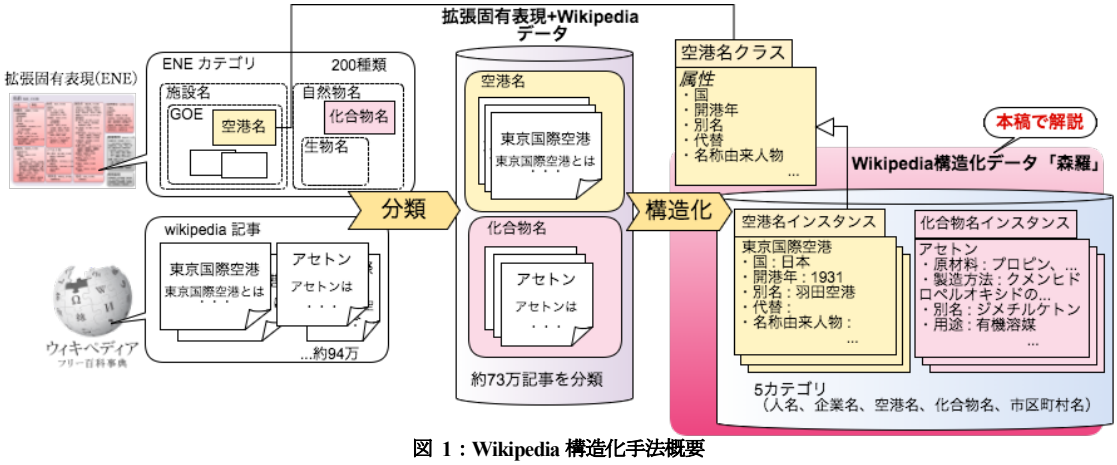

図 1：Wikipedia 構造化手法概要

で200種類の拡張固有表現が定義されている。［ENE definition HP］

3. 関連データおよび関連研究

構造化された知識ベースは自然言語処理全般において非常に重要な知識リソースと認識されている。過去においてこの問題に取り組んだ大型プロジェクトがいくつか存在する。古くはCYCプロジェクトから、最近ではWikipediaをベースにしたDBpedia、Yago、Freebase、Wikidataなどのプロジェクトである。また、共有タスクのプロジェクトとして知識ベースの構造化を目的としたKBPやFIGERといったプロジェクトもある。これらのリソースやプロジェクトについてここで紹介し、それらのプロジェクトにおいて我々が解決すべき課題と考えている点を述べる。

CYCプロジェクトは常識推論の実現を目指して作成された大規模知識ベースである[Lenat 95]。汎用ドメインの知識ベースは、人手で作られているため作成や保守のコストが非常に大きなものになっており、カバレージの点でも人手作成による限界が存在する。

DBpediaは、インフォボックスや上位下位関係知識などWikipedia内で半構造化されている情報を元に作られた構造化された知識である[Lehmann et al. 15]。このため、精度、カバレージ、一貫性などに問題がある。例えば、「新宿駅」は「小田急線」の下位概念として定義されているが、もちろん、駅は鉄道会社の下位概念ではない。元々の「新宿駅」のインフォボックスに属性が3種類しか値が設定されておらず、カバレージ低下の原因となっている。

YagoはWikipediaの項目をWordNetのオントロジーにマッピングすることによって作成されたオントロジーである[Mashdisoltani et al. 14]。WordNetは属性が定義されておらず、その部分はDBpedia同様にインフォボックスをそのまま利用しているため、DBpediaと同様、カバレージなどの問題がある。

FreebaseはWikipediaのようにクラウドソーシングによって構造化された知識ベースを作ろうという試みであった[Bollacker et al. 08]。しかし、手法からくる問題としてのノイズや一貫性のなさが各所に現れていて、一部のデータベースの複製である部分を除くと綺麗な知識ベースとは言えない。現在は以下に述べるWikidataプロジェクトに統合されている。

Wikidataは主にWikipediaの項目に対して構造化されたデータベースを作ることを目的としている[Vrandečić and Krötzsch 14]。Freebase同様にボトムアップに作成されているため、ノイズと一貫性の欠如の問題がある。

KBPはNISTによる共有タスクであり、構造化されてない文書から構造化された知識を抽出する技術を確立することを目標としている[KBP 17]。主要なタスクとしては２種類ある。文書中からそこで言及されている項目を見つけ出しDBエントリーを同定するタスク（EDL ：Entity Discovery and Linking）と、対象項目の属性値を抽出するタスク（SF ：Slot Filling）である。現状では対象項目のタイプは人名、組織名、場所名に限定されており、Wikipediaなどの幅広いタイプの項目をカバーするものではない。

FIGERは拡張固有表現のように、細かく定義された、112種類の固有表現を文書中から同定する共有タスクである[Ling and Weld 12]。構造化については扱われていない。

4. 分類結果の概要

分類作業とデータに関しては、[関根ら 18]に詳しいがここで簡単に説明する。Wikipediaの個別エンティティを表している項目が約73項目あり、それらを200種類の拡張固有表現に分類した。その過程は、まず人手で分類した約2万項目をトレーニングデータに機械学習で自動分類した。そして、その結果、怪しいと考えた項目については人手でチェックするという方法をとった。高頻度な拡張固有表現は表１にあげる。ここで、まず分類を行ったのは、構造化をするための前段階であり、構造化作業の効率化のためである。同じカテゴリの項目は同じ属性値を持つため、それぞれのカテゴリである程度のトレーニングデータを作成し、それを基に残りのデータを機械学習で構造化していこうと考えている。

表１．高頻度の拡張固有表現分類項目数

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

5. サンプルデータの作成

構造化のトレーニングデータを作成するために、以下の５種類のカテゴリにある項目を最低１４０項目ずつ取り出し、構造化作業を３種類の方法で試みた。

５種類のカテゴリ：

人名、企業名、市区町村名、空港名、化合物名

３種類の方法：

A) 言語データ作成経験者自身
B) 言語データ作成経験者の管理の下で学生
C) クラウドソーシング（ランサーズ）

これらの手法のうち、1,2については、言語データ作成経験者が直接・あるいは間接的に作業を行なうことで、直接Wikipediaから属性値の抽出を行っている。3については、クラウドソーシングの特徴を活かし、網羅率を向上させるための手法を考案した。

5.1. クラウドソーシングによるサンプルデータ作成

属性値の網羅率の高いサンプルデータを低コストに構築するために、クラウドソーシングサービス（ランサーズ）を利用した構造化作業を行った。クラウドソーシングでは、不特定多数の、知識、真面目さなどのバックグラウンドや指示の把握状況が不明な作業者に対して作業を依頼することになる。このため、クラウドソーシングでは、作業結果の品質を保証する枠組みを適切に設定することが重要となる。今回の作業では、次の二点に着目して作業タスクの設定を行った。

・タスク分割による指示の単純化
・タスクの冗長化・多段化

タスク分割については、どのような作業者であっても簡単に作業をできるように、構造化タスクを発見→抽出→検証の 3 段階に分割した。各ステップの概要を図 2 に示す。クラウドソーシングでは、言語データ作成経験者が介在する他の手法とは異なり、図に示すように、Wikipedia 記事から直接属性値を抽出することはせず、段階的に属性値を特定できるように作業を分割している。また、これらのステップそれぞれについて、タスクを冗長化・多段化することにより、品質が保証されるようにステップを設計した。以下、各ステップについてそれぞれ説明する。

発見：

発見ステップでは、作業者に、該当する属性値が書かれていそうな段落の抽出作業を依頼する。このとき、Wikipedia 記事が幾つかの領域に分割された状態で提示され、作業者は、その中から適切と思う領域を 1 つ選択する。この作業を順次異なる作業者に未選択の領域だけを残して提示しつづけ、2 名の作業者が属性値が含まれている領域が残っていないと回答した場合作業を終了する。この多段化により、属性値が含まれている可能性のある領域の取りこぼしが低減される。

抽出：

抽出ステップでは、発見ステップで検出された各領域から、実際に属性値を抽出する作業を依頼する。こちらも、発見ステップ同様、既に抽出された値を抽出済みとして明示し、他に属性値候補が残っていないかを順次異なる作業者に提示しつづける。このステップ

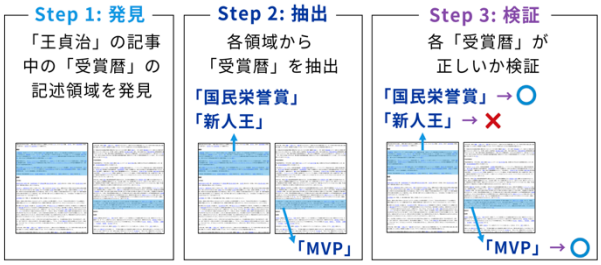

図 2：多段階クラウドソーシングによる構造化概要

も、2 名の作業者がこれ以上属性値が含まれていないと回答するまで作業を依頼し続ける。

検証：

検証ステップでは、抽出ステップで抽出された属性値候補が適切か否かを作業者に判断してもらう。このステップでは、それぞれの属性値候補について、二人の作業者に適切か否かを判断してもらい（冗長化）、一人でも適切であると判断した場合には正解とする。

5.2. サンプルデータの作成結果と考察

サンプルデータの構築結果グラフを図 3 に示す。結果より、クラウドソーシング(C)が最も再現率（マイクロ平均）が高くなっており、タスクの多段化・冗長化が有効に働いていると考えられる。一方で、精度も他と比べて数%差程度であり、タスク分割も有効であったと考えられる。特に、代表作や経歴など、作業者数に比例して網羅率が向上しやすい属性については、その効果が顕著に現れており、数十%高い結果となっていた。一方で、父親の名前など、ページ中の値すべての網羅が重要ではない属性については、言語データ作成経験者が介在する手法の方が高精度であった。

6. 問題点とその考察

構造化サンプルデータ作成に際して、いくつかの問題点が露見した。ここにそれらを列挙する。

構造化タスクのための属性定義変更の検討

構造化タスクにおいて、現状の拡張固有表現の属性定義では、あまり適切でないと考えられるものが有ることがわかった。例えば、人名の代表作について、作品のうちのどれを代表的なものと判断するのか、更には、参加したイベントや、発明した事物を指すエンテ

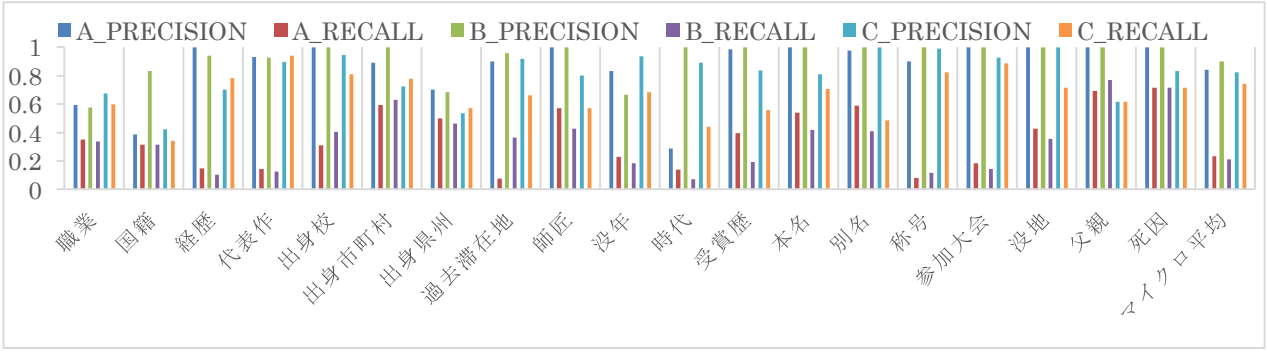

図 3：構造化アノテーション結果比較

ィティなどは人物にとっては代表作であるべきではないかと考えられる。一部、属性値が複数の属性に対応し得る（市区町村の「著名施設」と「史跡」など）ため、アノテーターによって結果がバラバラになるといった問題も発生したため、これらの変更、統合について検討する必要がある。

推測される属性値の扱い

上述の属性定義の問題に付随して、属性値がページ中に未記載であるものの、作業者にとっては推測が容易である場合についても検討する必要がある。例えば、出身市区町村名として、「台東区」などと記載されている場合に、出身都道府県についての記載がなくとも、その値が「東京都」であることは、容易に推測できる。今回はあくまで明記されている場合に限定して作業を行ってもらったが、その判断が難しいケースも存在すると考えられる（例えば、経歴に幼少期に東京で○○をしていた、などと記載がある場合など）ため、属性定義に値の抽出に関する指標の定義を付加する必要がある。

属性値の表記範囲に関する問題

記事内に非常に大量の属性値が記載されている場合（市区町村名の市区町村属性など）などに、作業者が属性値として抽出すべきであると正しく判断できないという問題が発生した。今回は、表記範囲すべてを取得することとしたが、人手による作業の限界を考慮した上で作業タスクの設定を検討する必要がある。

関係性ラベルの問題

属性の関係は時に単純な一つのラベルだけで表現することが困難な場合がある。例えば、ここまでに述べた時間の問題などに関連して、人名の「職業」などは一定期間のロールであるため、適切に表現することを考えた場合、属性自体を構造化して期間などの属性を付与できるように変更するべきか検討する必要がある。今回は、定義域にエンティティ名、値域に属性値となるエンティティ名を持った、一つのラベルで表現された属性を関係とするトリプレットで表現することとし、属性の構造化自体は見送ることとした。

7. Resource by Collaborative Contribution

全データの構造化を人手で行うことはほぼ不可能に近い。特に、日々更新されるWikipediaを対象にしているため、将来の更新作業を考慮しても現実的ではない。属性値抽出は機械学習によってある程度自動化できることが分かっており、今回のリソース作成でも機械学習を活用したいと思っている。現在の自然言語処理では「評価型ワークショップ」が多数行われている。この形式のワークショップはシステムの最適化競争の面があるが、これを利用して構造化データを作成することを計画している。つまり、運営者側で訓練データとテストデータを用意し、多くのシステムに評価型ワークショップに参加していただき、訓練データ以外の全項目を構造化することを考えている。この際にアンサンブル学習の手法を用いて、信頼できる出力を集めて自動的にリソースを作る。また、信頼度の薄いものを人手で確認訂正して次の学習の訓練データにするアクティブ・ラーニングや、何度も訓練データの作成とシステムの実行を繰り返すブートストラッピング手法を行うことで、多くの参加者を取り込みつつ、同時に精度の高いリソース作成を実現していくことを考えている。スケジュールは以下の通りである。

2017年12月6日：キックオフミーティング
2018年4月：　トレーニングデータ公開
2018年9月：　評価
2018年10月：　ワークショップ

この分野の多くの研究者の参加によって、本データがが構築されていくことを期待している。

8. まとめ

Wikipediaの構造化データ「森羅」の作成を目指したプロジェクトを推進している。前章に記した通りこのプロジェクトは多くの方の協力なしには進まない。より深い知識処理を実現するためにも、本プロジェクトに多くの協力をいただけるようにお願いしたい。

参考文献

[関根ら 18] 関根聡,安藤まや,小林暁雄,松田耕史,Duc Nguyen,鈴木正敏,乾健太郎 「拡張固有表表現+Wikipedia」データ（2015年11月版Wikipedia分類作業完成版）. 言語処理学会第42回年次大会(2018)

[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.

[Lenat 95] Douglas B. Lenat. CYC: a large-scale investment in knowledge infrastructure. ACM 38, pp. 32-38.

[Mashdisoltani et al. 14] Farzaneh Mahdisoltani, Joanna Biega, Fabian M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. Proceedings of the Conference on Innovative Data Systems Research (CIDR 2015).

[Lehmann et al. 15] Lehmann, J., Isele, R., Jakob, M., Jentzch, M., Kontokostas, D., Mendes, P.N., Hellman, S., Morsey M., Kleef, P., Auer, S. and Bizer, C. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. Semantic Web Journal, 6(2) :167-195

[Bollacker et al. 08] Bollacker, K., Evans, C., Paritosh, P., Sturge, T. and Taylor, J. Freebase: a collaboratively created graph database for structuring human knowledge. Proc. International conference on Management of data (SIGMOD '08). ACM, pp.1247-1250.

[Vrandečić and Krötzsch 14] Vrandečić, D. and Krötzsch, M. Wikidata: a free collaborative knowledgebase. Commun. ACM57, pp. 78-85.

[KBP 17] National Institute of Standards and Technology. Text Analysis Conference (TAC) 2017. https://tac.nist.gov/2017/

[Ling and Weld 12] Ling, X. and Weld, D.S. Fine-grained entity recognition. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12). pp.94-100.