Wikipedia構造化プロジェクト「森羅2018」

関根聡[1)] 小林暁雄[1)] 安藤まや[2)]
1) 理研AIP 2) ランゲージ・クラフト
{satoshi.sekine, akio.kobayashi}@riken.jp , ando@languagecraft.com

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 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に基づき、Wikipedia の説明文やインフォボックスから拡張固有表現で定義された属性に対する値を抽出した Wikipedia の構造化データーを作成することが目標である。本プロジェクトは、“Resource by Collaborative Contribution”の考えに基づき、多くの参加者により協力してリソースを作成していく形式をとっており、今回は 8 団体から 15 システムが提出された。この論文では、本プロジェクトの概要を説明し、その有効性を解説した上で森羅 2019 の計画についても紹介する。

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ある。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が利用可能な Wikipedia の構造化を進めている(図 1)。本稿では、[関根ら 18a]にて提案した「森羅 2018」プロジェクトについて説明する。

2. Resource by Collaborative Contribution

Wikipedia の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。しかし、属性値抽出は様々な機械学習手法によってある程度の精度で自動化できることが分かっている。今回の Wikipedia 構造化でも機械学習を活用するが、一つの機械学習システムだけで実現するのではなく、多くの違った種類のシステムが協力することによってより良いリソースを作成することを目標としている。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既存タスクにおける機械学習システムの最適化競争の側面があるが、これを逆に利用して構造化データを作成していこうと考えている。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く。この時にテストデータを参加者には知らせないことで、参加者には訓練データ以外の全項目を構造化するという仕組みを取り入れ、その結果は共有することを約束してもらう。この結果を利用しアンサンブル学習の手法を用いて、より信頼できるリソースを自動的に作る。また、信頼度の低いものを人手で確認訂正して次の学習時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取り入れることで、多くの参加者と協力しあって、精度の高いリソース作成を実現していくことを目標としている。

3. 森羅 2018 プロジェクトの実施

森羅 2018 プロジェクトは以下のスケジュールで実施され、8 団体から 15 システムの結果が提出された。

2017 年 12 月 6 日：キックオフミーティング
2018 年 4 月 24 日：開始説明会
　　トレーニングデータ公開
2018 年 9 月 10 日：結果提出、評価
2018 年 10 月 18：結果報告会

4. 森羅 2018 のタスク

今回のプロジェクトのタスクは、拡張固有表現の以下の 5 つのカテゴリーに属する項目から規定された属性値を抽出するタスクである。

〈対象カテゴリー〉
人名、企業名、市区町村名、空港名、化合物名

Wikipedia は 2017 年 11 月のバージョンを対象としており、転送ページ、リストページ、曖昧性回避ページを除く一般の項目の内、被リンク数 5 以上のポピュラーな 835,311 項目が今回のタスクの基になっている。これらを機械学習および人手チェックして分類したデータから、5 つのカテゴリーに分類されたデータが対象データとなる[関根ら 18b]。この内、それぞれのカテゴリーで属性値をアノテーターにより抽出した 600 項目をトレーニングデータとして公開し、100 項目をテストデータとして使った。ただし、前述の通りどのデータがテストデータとして使われたかという情報は公開していない。

構造化のフレームワークとしては広く自然言語処理応用を考えて名前空間のオントロジーとして定義された拡張固有表現を利用している。これは[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、芸術作品名などの幅広い固有表現や、地名以下には山地名、河川名、海洋名などの地形名や星座名などの天体名などが含まれる。拡張固有表現の Version 7.1.1 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。また、多くのカテゴリーにはそのカテゴリーに属する Wikipedia 項目の記載を基に、人手で作成した属性セットが定義されている。例えば空港名の属性では、国、母都心、年間利用者数、別名、名前由来人物などの属性が定義されている [ENE HP]。今回のプロジェクトでは、前述した 5 つのカテゴリーにおいて、Wikipedia 記事から定義された属性の値を抽出することで構造化された知識を作ることが目標である。

5. 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の確立を目標としている[KBP 17]。主要なタスクは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。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 等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。

これらの分析から、構造化される情報のカテゴリーや属性セットなどのフレームワークはトップダウンに設計する必要性がわかる。そして、その情報の内容はボトムアップに作成するか、ボトムアップに作成されたものを構造化していくかの方法が考えられる。

チーム	手法	人名	企業名	市区町村名	空港名	化合物名
TUT	人手作成パターン	20	41	28	**72**	
OCU	人手作成パターン	19				
NUT	機械学習(LightGBM)+パターン					42
SunSun	人手作成パターン		30			
Fuji Xerox	深層学習	31		43	42	39
Toppan	パターン＋深層学習		33		35	
Unisys	DrQA（質問応答ツール）	**44**	**53**	42	67	**47**
AIP	深層学習	36	38	**46**	71	46

表１．参加システムと評価結果

6. 参加システムと結果

今回の「森羅 2018」プロジェクトには 8 団体から 15 システムが参加した。各参加者の代表的なシステムが用いた手法と、各カテゴリーの評価結果の F 値を表 1 に載せる。それぞれの参加チームが全てのカテゴリーで参加したものとは限らない。またこの表では締め切りを過ぎてから提出した結果も含まれている。

全体的には、空港名では 70 近い F 値萌えられているが、それ以外では最高で 50 程度の結果しか得られておらず、基本的に非常に難しいタスクであったことが伺える。手法については、人手によって正規表現のようなパターンを作る方法か深層学習を使う方法がポピュラーである。DrQA は質問応答タスク向けに公開されているツールで、日本語化したものを利用している。このシステムはある特定のカテゴリーの属性値を求める場合には、それを自然文の質問にして与えたり、インフォボックスのような表に入っているデーターも自然文に直したデータを作成してから DrQA を利用している。

基本的には DrQA のシステムが全体的に優れた結果を残している。詳細な分析はシステムを持ってないので難しいが、これはトレーニングデータが少ないため、RdQA のシステムがカテゴリーや属性に横断的に学習を行い、より大きいトレーニングデータを使って学習できたためという効果があるという見方ができる。また、空港名の結果が良かったのは、インフォボックスにある情報が属性情報と数多く重なっていた上、インフォボックスのテンプレートがほぼ 1 つしかなく、人手で作成したパターンでもこのような情報が十分に獲得できたということが挙げられる。化合物名については、インフォボックスにあるような情報が役立ったのと同時に、「用途」「製造方法」などの長い表現が属性値となるような属性もあり、そのようなものは非常に精度が低かったことが挙げられる。また、人名、企業名、市区町村名では、代表作品、観光地などの複数の値のリストになるような属性は精度が低いことが観察されている。このような値は、パターンでも深層学習でも難しいことが考えられる。パターンベースのシステムがあまり高くないのは、インフォボックスのテンプレートの種類が、人の職業によってや企業の業種によって異なっており、それらを網羅することが厳しかったのではないかと考えられる。また、評価方法に関する大きな問題としては、正解と出力結果の部分一致の問題が見つかっている。シンプルな記号の過不足などがあるが、システムの出力の形でも正解とも考えられるが正解とは異なっている場合には不正解となってしまうために過度に制度が低くなっている問題である。MediaWiki 記法、またそれを変換した HTML タグの記述によって抽出が難しいものと、本文中の属性を表す表現の訓練データ中での出現頻度や文中での表記の差異によって正しく抽出されていないといった問題が見つかっている。より詳細な結果報告やエラー分析については(小林ら 19)を参照していただきたい。

7. アンサンブル学習

RbCC の考え方が実際に有効に働くかどうかは、全システムの結果を用いてアンサンブル学習を実施してみた際に、より良い結果が得られるかどうかで調べることができる。この詳細な報告は(中山 19)で報告するが、本論文では表２に結果を示し、概要だけを説明する。表２では、各カテゴリー毎の最も良いシステムの F 値、アンサンブル学習を行った後の統合システムの F 値、その向上率を載せてある。結論からすると、人名、空港名、化合物名の３つのカテゴリーにおいてアンサンブル学習にて F 値が 10 以上上昇するという非常に好ましい結果が得られた。他の 2 つのカテゴリーにおいてもポジティブな結果が見られ、アンサンブル学習、ひいては、RbCC の考え方が非常に有効であることが証明された。特に今回は人手によるパターンのシステム、深層学習のシステム、DrQA のシステムと非常に異なった手法に基づくシステムが提出されており、その理由によりアンサンブル結果が良かったとも考えられる。別の見方をすると、（精度、再現率ともに上がっている）この結果は深層学習では拾えない正解データが存在すること、深層学習で間違えた答えをパターンベースのシステムや別の手法で排除できることを意味し、今後のシステム開発の方針にも重要な示唆を与えていると考えられる。

カテゴリー	最高システム	アンサンブル結果	向上値
人名	43.9	47.9	+4.0
企業名	53.4	60.8	+7.4
市区町村名	46.0	58.4	+12.4
空港名	71.8	87.0	+15.2
化合物名	47.1	64.8	+17.7
平均	51.4	62.7	+11.3

表２．アンサンブル学習結果（F 値）

8. 「森羅 2019」およびデータ公開

Wikipedia の構造化データ「森羅 2018」は当初の目的を達成し、RbCC の考え方の有効性が実証されたと考えている。この考え方を推し進め「森羅 2019」において以下の３つのタスクを実施しようと考えている。

ML：多言語化（９ヶ国語の分類）
英語、スペイン語、フランス語、ドイツ語、中国語、ロシア語、ポルトガル語、イタリア語、アラビア語の９カ国後に対して、日本語からのリンクを教師データにした分類タスク

JP-5:「５カテゴリー」
より大きなアノテーションによるトレーニングデータを用意し 2018 と同じ５カテゴリーで行う構造化タスク

JP-34：幅広いカテゴリー
お互いに類似した 34 カテゴリーにおいて各 100 項目のトレーニングで行う構造化タスク

3 月 12 日から始まる言語処理学会の年次大会時には JP-5, JP-34 のデータの準備が完了し、タスクが始まっている予定である。日程的には昨年同様に、説明会を 4 月に、結果提出を 9 月に、結果報告会を 10 月頃に実施する予定である。

また同時に、Wikipedia 項目の拡張固有表現を基にした分類データを一般公開する計画である。森羅 2019 の JP-5, JP-34 のタスクでは 2017 年 11 月の Wikipedia のダンプデータを利用するが、ML タスクでは 2018 年 11 月またはそれ以降の新しい Wikipedia のダンプデータを利用する。この ML タスクの元データとなる日本語の Wikiepdia の分類データを 2019 年の 4 月頃に一般公開する予定にしている。

9. 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。「森羅 2018」に協力いただいた皆様、特に評価に参加いただいた８団体にはここで感謝を述べたい。今後も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, 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 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24回年次大会(2018)

[小林ら 19] 小林暁雄, 中山功太, 関根聡「森羅:Wikipedia 構造化プロジェクト 2018 結果の分析と考察」. 言語処理学会第 25 回年次大会(2019)

[中山ら 19] 中山功太, 小林暁雄, 関根聡「共有タスクにおける GA 重み付け加重投票を用いた属性値アンサンブル」. 言語処理学会第 25 回年次大会(2019)

[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)

[ENE HP]https://sites.google.com/site/extendednamedentity711/

[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.