公開

研 究 終 了 報 告 書

「31 言語における FG-NER・EL システム開発」

研究期間: 2020 年 4 月～2023 年 4 月
研 究 者: 中山功太

1. 研究のねらい

近年、ウェブの発達に大規模な情報をより簡単に得ることができるようになった。しかし、これらの情報は主に自然言語で記述されており、コンピューターで扱うことに適した形式ではない。AI 技術の発展により、自然言語を直接入力にとり、様々な予測を行うことが可能になったが、その推論過程のブラックボックス化が問題となっている。本研究の目的は、自然文を固有表現レベルで構造化することであり、これは AI による推論過程の可視化を行うための根幹技術となる。

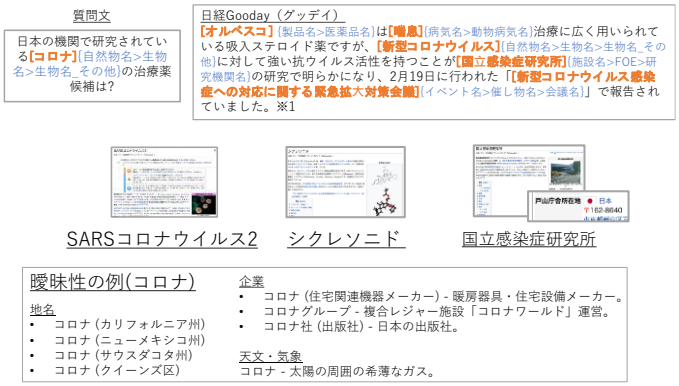

図 1 細分類固有表現抽出・エンティティリンキング理想出力

本研究では主に、細分類固有表現抽出(FG-NER)とエンティティリンキング(EL)システムの開発を行う。理想とするシステムの出力結果を図 1 に示す。この例では、質疑応答システムの想定入力に対して、細分類固有表現抽出とエンティティリンキングを適用している。固有表現抽出では、文章中の固有表現を検出し、固有表現クラスに分類する。本研究ではクラス定義に拡張固有表現階層を使用しており、200 近いクラスに分類を行う。エンティティリンキングでは、固有表現を事前定義されたエンティティリストへ紐づける。本研究では Wikipedia の各記事に対して紐付けを行う。以上により、固有表現レベルで情報を構造化でき、AI はこれらの情報を推論根拠の一つとして提示できるようになる。

細分類固有表現抽出とエンティティリンキングシステムを学習するためには、教師データが必要であるが、作成に膨大なコストがかかる。本研究では多言語適用も考えており現実的ではない。そのため、Wikipedia の内部リンクを活用してシステムを学習することを考える。しかし、Wikipedia の編集ガイドラインにより、自己記事にリンクする固有表現に関してはリンクしない、複数回出現した固有表現に関しては最初の出現のみリンクするといったリンク付与ルールが定められている。そのため、教師データとして使用するにはラベルが不完全である。また、当然ではあるが Wikipedia 上に存在しない記事に対してもリンクは付与されない。本研究では、これらの不完全なラベルから言語非依存にシステムを学習する方法を考案する。また、実際に日本語だけでなく、多言語に対するシステム構築を行う。以上により、AI は言語を跨いで推論根拠を提示することができ、より信頼される AI に関する研究が発達すると考える。

公開

2. 研究成果

(1)概要

米 ツイッター{チャンネル名}(Twitter) の買収意向を示している米電気自動車大手 テスラ{企業名}(テスラ(会社)) の イーロン・マスク{人名}(イーロン・マスク) 最高経営責任者(CEO){地位職業名}(最高経営責任者) が最近、ロシア{国名}(ロシア) や 中国{国名}(中華人民共和国) 寄りと受け取れる言動を繰り返し、懸念が強まっている。両国による 情報操作{事故事件名__その他}(情報操作) の問題が指摘されるなか、マスク氏{人名}(イーロン・マスク) が買収した場合の ツイッター{チャンネル名}(Twitter) の投稿管理に不安の声が上がる。[引用: https://news.yahoo.co.jp/articles/13ec34b749d7688241db4593d0505795bea4d346]

NTT都市開発{企業名}(NTT 都市開発) と シンガポール{国名}(シンガポール) のホテルチェーン、カペラホテルグループ{企業名}(NIL) は12日、大阪城公園{公園名}(大阪城公園) と 難波宮跡(なにわのみやあと)公園{公園名}(難波宮) の間に位置する NTT西日本旧本社{企業名}(西日本電信電話) 跡地に高級ホテル「 パティーナ大阪{宿泊施設名}(NIL) 」(大阪市{市区町村名}(大阪市) 中央区{市区町村名}(中央区(大阪市)))を令和7年春に開業すると発表した。
[引用: https://news.yahoo.co.jp/articles/2797b755d1f58a7d03ad4c6690789d34f4edac2c]

図 2 システム実行結果

ACT-X 期間中における最も重要な成果は「Wikipedia を活用した細分類固有表現抽出・エンンティティリンキンシステムの学習方法の考案」である。考案した手法をもとに実際に日本語の Wikipedia を用いてシステムを学習し、WEB ニュースに対して予測を行った結果を図 2 に示す。ここで青文字下線部はテキスト上から検出された固有表現の言及を示す。{緑文字波括弧内}は細分類固有表現抽出の結果であり、固有表現が属する固有表現クラスを示す。この固有表現クラスは、拡張固有表現階層で定義されたものを採用している。(オレンジ文字括弧内)はエンティティリンキングの結果であり、固有表現に対応する Wikipedia のページを示す。(NIL)は対応するページが存在しないことを意味する。

図 2 の予測結果を見ると、上文章中の「Twitter」や「イーロン・マスク」を指す固有表現は文章中に複数回登場するが、再度出現した固有表現に関しても予測できている。また、下文中の「カペラホテルグループ」や「パティーナ大阪」は Wikipedia に存在しないため、モデルは(NIL)を出力しており、Wikipedia に存在しない未知の固有表現も検出できている。これらの結果は編集ガイドライン等によるラベルの欠落を提案手法により解決できていることを示す。また、上文章中の「テスラ」は、①{人物名}(ニコラ・テスラ)②{単位名__その他}(テスラ(単位))③{企業名}(テスラ(会社))④{惑星__衛星名}(テスラ(小惑星))等の候補があり、エンティティリンキングにおいて非常に難しい問題である。システムは、正答の③{企業名}(テスラ(会社))を出力しており、文脈を加味した予測ができていることが窺える。

また、本課題では日本語以外の 30 言語におけるエンティティリンキング・再分類固有表現抽出システムの学習も目標としている。実現のためには、Wikipedia 記事が固有表現クラスに分類されている必要がある。日本語はすでに人手によって分類されているため、このデータをもとに多言語適用が行われた。具体的には Wikipedia の構造化を目指す森羅プロジェクトにおいて、複数の参加者を募りタスクを解くといった共有タスクの形式で分類が行われている。しかし、複数参加者から結果が提出されており、実際に使用するためにはこの結果に対し統合を行う必要がある。一般的に予測結果の統合を行うためには専用の正答データが必要であるが、30 言語で本データを作成することは非常にコストがかかり現実的ではない。そのため、専用の正答データを必要としない統合手法を考案した[1]。本成果により、日本語以外の 30 言語の分類結果が得られ、冒頭の手法と組み合わせることで本課題が実現可能となる。多言語適用後の結果も国際学会等に投稿予定である。

公開

（2）詳細

研究テーマ A「共有タスクにおける提出結果の統合手法開発」

Wikipedia から固有表現抽出用の教師データを作成するためには、Wikipedia 記事は事前に固有表現クラスに分類されている必要がある。本研究では図 3 に示す約 200 クラスを対象としている。日本語 Wikipedia は既に人手で分類されているが、多言語適用を考えた場合、他言語の Wikipedia も分類する必要がある。Wikipedia では、言語を跨いで同一のエンティティが存在した場合、言語間リンクで紐づけられている。そのため、他言語においても日本語からリンクしている記事を教師ラベルとして扱うことで分類が可能となる。Wikipedia の構造化を目的とする森羅プロジェクトがこの多言語分類を行なっており、参加者を募り共同で解くといった共有タスクの形で行われている。参加者はそれぞれが開発した分類システムの結果を提出しているため、本研究で用いるためにはこれらの結果を統合する必要がある。一般的に精度のよい統合を行うためには学習に使用されていない統合用の正答ラベルが必要であるが、30 言語でこれらのデータを用意することは非常に困難である。

図 3 拡張固有表現階層

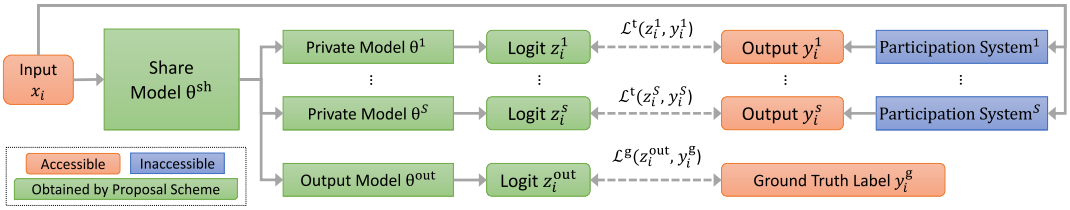

図 4 共同教育(Co-Teaching)

本テーマでは図 4 に示す共同教育(Co-Teaching)という統合手法を提案した。共同教育では、共有タスクに提出されたシステム結果を教師データとして、単一の深層学習モデルを学習することで統合を行う。この際、共有タスクで使用した学習データも教師データとして使用する。本手法は、複数の教師が一人の生徒を教えているような形をとるため、共同教育と呼ぶ。本手法の利点は、統合専用の正答ラベルを必要としない点に加えて、システム自体を統合可能であるため、新規の入力データや入力データに更新があった場合にも対応できる点である。日本語属性値抽出タスクの結果に対して統合と評価を行い、ベースライン手法より高い精度で統合ができていることを示した。本研究成果は EMNLP Findings に採択済みである。

公開

研究テーマ B「Wikipedia を活用した細分類固有表現抽出・エンンティティリンキングシステムの学習方法の考案」

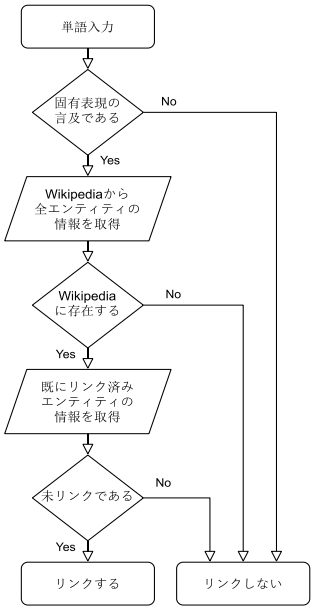

図 5 リンク付与フローチャート

研究のねらいで述べたとおり、細分類固有表現抽出及びエンティティリンキングシステムの学習データの作成は非常にコストが掛かる。そのため本研究テーマでは、Wikipedia の内部リンクを活用して、システムの学習を行う。しかし、これらは編集ガイドラインや Wikipedia に存在しないエンティティの影響により、リンクが大きく欠損している。そのため、本テーマでは、この欠損したリンクの影響を最小限にとどめながらシステムを学習する手法を考案した。

提案手法では、Wikipedia のリンク付与ルール自体を深層学習で模倣し、固有表現を検出する部分のみをモデルから取り出すことで欠損したラベルの影響軽減を行なっている。具体的には、Wikipedia の編集者は図 5 に示すようなフローチャートに従いリンクの付与をおこなっていると考え、これを再現可能な深層学習モデルを設計した。

モデル構造詳細に関しては未発表のため省略する。

図 2 に本システムの予測結果が示されている。上文章中の「Twitter」や「イーロン・マスク」を指す固有表現は文章中に複数回登場するが、再度出現した固有表現に関しても予測できている。また、下文中の「カペラホテルグループ」や「パティーナ大阪」は Wikipedia に存在しないため、モデルは(NIL)を出力しており、Wikipedia に存在しない未知の固有表現も検出できている。これらの結果は編集ガイドライン等によるラベルの欠落を提案手法により解決できていることを示す。また、上文章中の「テスラ」は、①{人物名}(ニコラ・テスラ)②{単位名__その他}(テスラ(単位))③{企業名}(テスラ(会社))④{惑星__衛星名}(テスラ(小惑星))等の候補があり、エンティティリンキングにおいて非常に難しい問題である。システムは、正答の③{企業名}(テスラ(会社))を出力しており、文脈を加味した予測ができていることが窺える。以上の結果は、評価データを作成し定量的に評価も行った上で、国際学会および学会誌に投稿予定である。

公開

3.　今後の展開

今後は ACT-X 期間中に開発した技術の社会実装のため、API 基盤の構築と知識ベース構築を行う。以上を 1 年以内に行う予定である。

API 基盤構築　ACT-X 期間中に構築したシステムを一般の方が利用しやすい形で提供するには

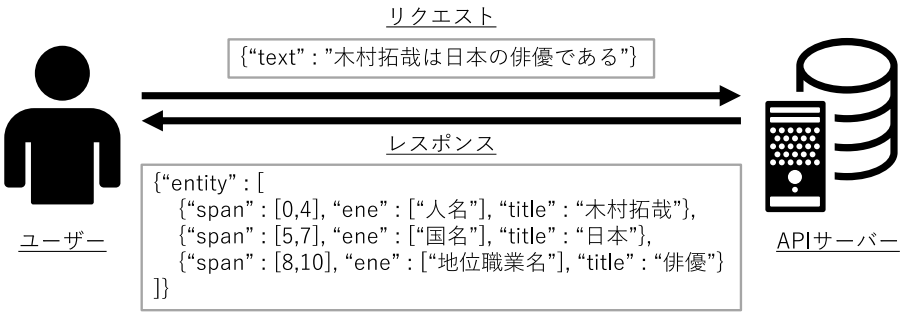

図 6 API サーバー概要

API 公開が最も現実的である。固有表現抽出とエンティティリンキングを同時に行える API はほとんどの言語で存在しないため、多くの需要があると考えられる。しかし、API サーバーを運用し続けるには維持費がかかるため、サーバー自体は公開せず、サーバーに容易にインストール可能な API 基盤のみを構築することとする。API 基盤は Docker コンテナ等の仮想環境上で構築することで、配布可能とする予定である。API サーバーのイメージを図 6 に示す。システムを使用したいユーザーは、サーバーに対して json 等の形式でリクエストを送る。サーバーは受け取った情報をもとに深層学習モデルによる計算を行い、json 形式でレスポンスをユーザーに返す。API 基盤の構築で問題となるのは主に深層学習モデル部分の計算コストである。ACT-X の成果で提案したモデルは非常にパラメーター数が多く、頻繁に API にアクセスが想定される場合、高性能な GPU が必要となる。API が広く使われるためには低価格帯の GPU、もしくは CPU のみで動くモデルが好ましいため、深層学習モデルの計算コストを削減する必要がある。その実現のために現在 2 つのアイディアがある。前者は知識蒸留であり、これは元の大規模な深層学習モデルの予測結果を学習データとしてよりパラメーター数の少ない深層学習モデルを学習することでパラメーター圧縮を行う手法である。後者は量子化であり、内積等の計算を近似することで計算コストの削減を行う手法である。以上により、実用に耐えうる API 基盤の作成を行う。

知識ベース構築　ウェブ社会において、インターネット上では、さまざまな情報が飛び交っている。これらの多くは自然言語で記述されており、コンピューターで扱うにはふさわしくない形式である。ACT-X において開発したエンティティリンキング・細分類固有表現抽出システムの結果をもとに、自然言語をコンピューターが扱いやすい知識ベースの形へと落とし込むことでより活用することができると考える。今回想定している知識ベースは、図 7 のように各ノードがエンティティを表し、各エッジが 1 つの Web 記事を

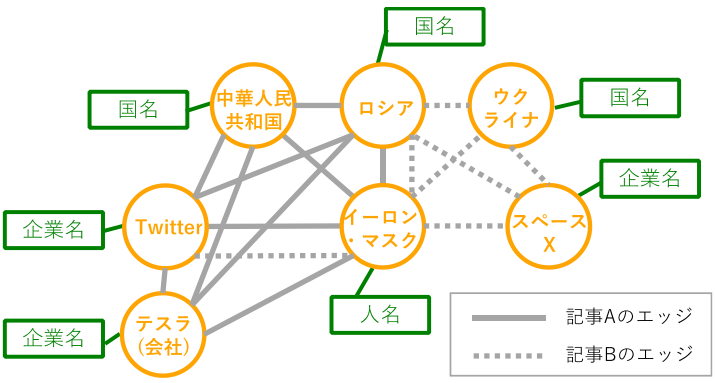

図 7　知識ベース概要

公開

表すものとする。各ノードは属性として固有表現クラスラベルを持っている。エッジは 2 つのエンティティが同一 Web 記事に出現したことを示している。また、言語を跨いで同じエンティティが存在する場合、同一エンティティであることを示すエッジも付与する。この知識ベース自体は、直接検索エンジンとして利用することが可能である。通常の検索エンジンと異なる点は、エンティティを指定して検索できるため、名前の揺れを考慮する必要がないところである。また、固有表現ラベルを使用して検索できるため、例えば〈イーロン・マスク〉と{企業名}で検索することで、イーロン・マスクといずれかの企業が関連する記事を得ることができる。これは検索エンジンだけでなく、金融技術やファクトチェックといった、多くの研究に適用可能である。実際に知識ベースを構築し、ダンプファイルの形式で公開する予定である。

4. 自己評価

研究目的の達成状況 システムを学習するための手法開発は終了しており、概ね当初の計画通り進行している。評価データの作成のみ遅れており、評価データのアノテーションを行うためのツール開発は終わっているものの、実際のアノテーションへは進めていない。本年度末までのアノテーション完了を目標に進めている。

研究の進め方（研究実施体制及び研究費執行状況） 概ね計画に沿った研究実施体制が取れており、エフォートに相当する研究時間が確保できていた。また、研究費も滞りなく執行できているが、アノテーションに関係する費用のみは現状未執行である。本年度末までに評価用データ作成のため、執行予定である。

研究成果の科学技術及び社会・経済への波及効果 領域のテーマの一つに「信頼される AI」とある。信頼される AI は、推論根拠を明示できることが必要不可欠である。本成果は、推論根拠を導出するための根幹技術であり、社会が直面する課題解決の大きな一歩となったと考えている。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：1件

1. Kouta Nakayama, Shuhei Kurita, Akio Kobayashi, Yukino Baba, and Satoshi Sekine. 2021. Co-Teaching Student-Model through Submission Results of Shared Task. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4525–4535, Punta Cana, Dominican Republic. Association for Computational Linguistics.

複数の参加者を募り共同でタスクを解く共有タスクでは、最終的に複数の結果が得られる場合が多い。本論文では予測結果を統合するための手法である共同教育(Co-Teaching)を提案している。本手法は、参加者の提出結果を教師データとし、共有タスクで使用された学習データと併せて、新たに深層学習モデルを学習することで、統合専用のラベル付きデータを必要とせず統合を行うことができる。実際に共有タスクの結果に対して統合を行いその性能を評価している。

（2）特許出願

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）