日本語 Wikification コーパスの構築に向けて

Davaajav Jargalsaikhan* 岡崎 直観† 松田 耕史† 乾 健太郎†
* 東北大学工学部 † 東北大学大学院情報科学研究科
{davaajav, okazaki, matsuda, inui}@ecei.tohoku.ac.jp

1 はじめに

エンティティ・リンキング（Entity Linking: EL）は、与えられた文章中のメンション（エンティティに言及する表現）を認識し、Wikipedia や Freebase などの知識ベースのエントリに対応付ける処理である。特に、Wikipedia を知識ベースに採用したエンティティ・リンキングは Wikification [8] と呼ばれる。例えば、

> サッカー の ワールドカップ（W杯）へ調整を続ける 日本代表 は 2 日、W杯 会場となる 神戸市 の 神戸ウイングスタジアム で ホンジュラス とのキリン・カップ第 2 戦 に臨む。

という文に対し、Wikification は「サッカー」を “サッカー”、「ワールドカップ」と「W杯」を “2002 FIFA ワールドカップ”、「日本代表」を “サッカー日本代表”、「神戸市」を “神戸市”、「神戸ウイングスタジアム」を “御崎公園球技場”、「ホンジュラス」を “サッカーホンジュラス代表”、「キリン・カップ第 2 戦」を NIL に、それぞれ対応づける [1]。

EL は質問応答 [5]、情報検索 [1]、知識ベース拡充（Knowledge Base Population）[2]、共参照解析 [3] など、知識に基づく言語処理を実現するために必須のタスクである。英語を対象とした EL では、UIUC 群（ACE と MSNBC）[9]、AIDA 群 [4]、TAC-KBP 群（2009 年から 2012 年まで）[7] など、10 件を超えるデータセットが存在する。対象とするエンティティや知識ベースの種類などのタスク設定が統一されずにデータセットや研究が乱立しており、研究の最先端が分かりにくくなっているという指摘すらある [6]。

一方で、日本語の EL に関する研究は少ない。古川らは、日本語の学術文献に出現する専門用語を英語の Wikipedia に対応付けた [14]。林らは、機械翻訳での応用を念頭に、日英対訳文で日本語と英語の Wikification を同時に行う手法を提案した [15]。中村らは、任意の言語で記述されたソーシャルメディアの投稿に含まれるキーワードを英語の Wikipedia に対応付け、言語横断的なトピック抽出を目指した [13]。長田らは、都道府県別のニュース配信を想定し、テキスト中の人物、組織、場所のエンティティを都道府県の粒度で対応付ける研究を発表した [12]。

残念ながら、これらの研究は日本語の知識ベースへの EL に重点を置いていない。2016 年 1 月現在、日本語の Wikipedia には約 100 万件の記事、英語の Wikipedia には約 500 万件の記事が収録されているが、日本語から英語への言語間リンクは約 56 万件しかない。知識ベースの多くは Wikipedia に由来しているため、英語の知識ベースが日本の国土、経済、文化に関するエンティティを網羅するとは到底期待できない。また、日本語のテキストで英語の知識ベースを拡充する際にも、日本語の EL システムが必要になる。

さらに、日本語の EL の諸問題は日本語の固有表現抽出と関連づけて分析するべきであるが、それを可能にするようなコーパスも存在しない。そこで、本研究では、関根の拡張固有表現階層 [2] に基づく固有表現がアノテーションされた BCCWJ の新聞記事に対し、日本語版 Wikipedia を知識ベースとした Wikification のデータセットを構築する。

2 データセットの構築

2.1 設計方針

Ling らが指摘するように、EL タスクを正確に定義することは難しい [6]。EL の対象を固有表現のみに限定するか、一般名詞も含めるか、対象とする固有表現の意味クラスの範囲、固有表現の境界の設定、エンティティの特定性、換喩（メトニミー）の取り扱い等、様々な取捨選択が必要である。

例えば、1 節の例文では、「ワールドカップ」というメンションに対して、“ワールドカップ”、“FIFA ワールドカップ”、“2002 FIFA ワールドカップ” 等がリンク先の候補になる。これらの候補はすべて正解なので、より特定性の高い後者のエントリを付与したくなる。しかし、Wikipedia がエンティティを均質に収録している保証はない。例えば、2016 年 1 月現在、「2034 年 FIFA ワールドカップ」という記事は存在しないため [3]、仮に 18 年後のワールドカップについて議論しているテキストでは、「ワールドカップ」を “FIFA ワールドカップ” とするのか NIL とするのか、悩むことになる。

[1] 本論文では、Wikipedia のエントリを二重引用符で囲む。また、NIL は Wikipedia のエントリに対応付けられないことを表す。

[2] https://sites.google.com/site/extendednamedentityhierarchy/

[3] 驚くべきことに、2016 年 1 月時点において「2030 FIFA ワールドカップ」という記事が存在している。

また、「神戸ウイングスタジアム」において、「神戸」を EL の対象に含めるかどうか、判断に迷うかもしれない。また、日本語 Wikipedia には “神戸ウイングスタジアム” という記事があるが、この記事によると「神戸ウイングスタジアム」は球技場の運営管理企業の名称であり、球技場の正式名称は “御崎公園球技場” と説明されている。また、「ホンジュラス」は “ホンジュラス”（国）ではなく、“サッカーホンジュラス代表” を意味する換喩である。

これらの問題は、EL の対象となる固有表現の意味クラスを特定しておくことで、ある程度解消される。とは言え、現時点では付与対象とする固有表現の意味クラスに関して、参考にできる指針もない。そこで、本研究では固有表現の意味クラスが付与されたコーパスを出発点とし、できるだけ広範囲にアノテーションを行うことで、日本語の EL タスクの設計方針そのものから検討することにした。本論文では、BCCWJ に収録されていて、関根の拡張固有表現階層により固有表現の意味クラスが予め付与されている 340 件の新聞記事のうち、274 件に付与した成果を報告する。

2.2 付与手順

エンティティを人手で効率よく付与するため、brat rapid annotation tool (brat) [4] を用いた。brat にはテキストを Wikipedia や Freebase などの外部知識ベースに関連付ける機能が実装されている。そこで、日本語 Wikipedia のスナップショットから、各記事の ID、タイトル、概要セクションのリード文（説明文）を抽出し、brat 用の知識ベースとした。

関根の拡張固有表現階層には、200 種類の意味クラスが定義されている。このうち、時間表現（12 種類）、数値表現（34 種類）、アドレス（5 種類）、称号名（2 種類）、施設部分名（1 種類）に属する固有表現を EL の対象から外し、それ以外の固有表現に対して Wikipedia 記事へのリンクを付与した。リンク可能な候補が複数ある場合は、できるだけ特定性の高い Wikipedia 記事を選ぶ。また、Wikipedia の曖昧性解消ページ、カテゴリーページ、WikiMedia ページへのリンクは禁止とした。さらに、Wikipedia 記事中のセクションに対応付けたい場合は、リンク先をそのセクションを含む記事とした。付与作業は 3 人の日本人大学生が担当したが、付与箇所に重なりがないため、現時点ではアノテーションの一致率を測定できない。

2.3 付与結果

表 1 に、本研究で構築したデータセットの統計情報を示した。2.2 節の基準により EL の対象となった 22,514 件のメンションのうち、16,526 個が日本語 Wikipedia 記事にリンクされた。5,689 種類のメンションが、5,435 種類のエンティティにリンクされている。

[4] http://brat.nlplab.org/

表 1: 構築したデータセットの統計情報

項目	値
記事数	274
メンション数	22,514
リンク数	16,526
NIL 数	5,988
メンションの種類数	5,689
エンティティの種類数	5,435

表 2 に、メンションが 100 回以上出現した拡張固有表現意味クラスのうち、エンティティが付与された割合が上位の 10 クラスと下位の 10 クラスを示す。Province（都道府県州名）や Pro_Sports_Organization（プロ競技組織名）のように、よく知られている地名やチーム名は Wikipedia に収録されることが多く、リンク率が高くなっている。一方、下位 10 件のうちの Corporation_Other（法人名_その他）、Conference（会議名）などに関しては、2.4 節で記述するような部分・全体、固有表現のネストが多かったため、リンク率が下がった。また、Person（人名）や Character（キャラクター名）はあまり一般的ではない固有表現も多く、Wikipedia に収録されていないことが多かったため、リンク率が低下した。

語義曖昧性解消タスクでは、談話内で同じメンションが取りうる意味は 1 つという仮定（one-sense-per-discourse）がよく用いられる。本データセットで調べた所、この仮定が成り立つのは 274 文書のうち 229 記事（83.58%）であった。この仮定が成立しない例として、「タイガース」が “デトロイト・タイガース” と “阪神タイガース”（プロ野球選手の大リーグ移籍)、「ブッシュ」が “ジョージ・H・W・ブッシュ” と “ジョージ・W・ブッシュ”（親子で、前者は「大ブッシュ」、後者は「小ブッシュ」と呼ばれることもある) の両方を指す事例があった。また、「東京」が “東京” と “東京テレビ”、「PHP」が “PHP 研究所” と “PHP 新書” など、換喩によって one-sense-per-discourse 仮定が崩れる事例もあった。

2.4 付与が難しい事例

付与作業を行った際、判断に迷う事例もあった。このようなケースは、場所や組織の部分・全体、固有表現のネスト、時間経過によるエンティティの変化の 3 つに大別できる。以下、これらの事例を紹介する。

部分・全体

> 昨年の噴火は 洞爺湖温泉街:Location_Other の背後に火口をつくり、その前は山塊の形を変え、さらにその前は昭和新山を生んだ。

表現「洞爺湖温泉街」は洞爺湖温泉の一部である街のこと指しており、洞爺湖温泉のことを指している訳ではない。Wikipedia では洞爺湖温泉の記事は存在するが、洞爺湖温泉街の記事ではないので、「洞爺湖温泉街」を NIL にした。

固有表現のネスト

表 2: 100 回以上出現した意味クラスのうち、エンティティが付与された割合が上位の 10 クラスと下位の 10 クラス

カテゴリー名 (日本語名)	例	リンク率	リンクされた数	出現数
Pro_Sports_Organization (プロ競技組織名)	読売ジャイアンツ	0.984	249	253
Province (都道府県州名)	福岡県	0.979	610	623
Country (国名)	アメリカ合衆国	0.973	1783	1832
GPE_Other (GPE_その他)	銀座	0.971	99	102
Political_Party (政党名)	自民党	0.951	213	224
City (都市名)	仙台市	0.932	1245	1336
Continental_Region (大陸地域名)	アジア	0.914	117	128
Mammal (哺乳類名)	カンガルー	0.883	158	179
International_Organization (国際組織名)	NATO	0.881	192	218
Company (会社名)	NTT	0.880	587	667
	...			
Person (人名)	鈴木宗男	0.592	2148	3624
Corporation_Other (法人名_その他)	宇宙開発事業団	0.562	253	450
Conference (会議名)	衆議院本会議	0.551	70	127
Public_Institution (公共機関名)	高槻市役所	0.502	108	215
Organization_Other (組織名_その他)	総務課	0.478	55	115
Political_Organization_Other (政治的組織名_その他)	竹下派	0.434	59	136
GOE_Other (GOE_その他)	ホワイトハウス	0.361	99	274
Plan (計画政策名)	所得倍増計画	0.258	29	112
Occasion_Other (催し物名_その他)	つくば科学万博	0.119	26	218
Character (キャラクター名)	ミッキーマウス	0.107	11	103

その役割は当初、直接選挙を主張するイスラム教シーア派最高権威:Position_Vocation、シスタニ師に断念を説得する一時的な火消し役と考えられていた。

「イスラム教シーア派最高権威」は特定の宗教における権威なので、“権威” に紐付けしにくい。一方、“シーア派” は対象表現とは関連があるが、権威でなく宗教を指しているので「イスラム教シーア派最高権威」を NIL にした。

時間経過によるエンティティの変化

伊藤代表はあいさつで、知事選: Event_Other や札幌市長選について、「出馬する候補の政策を検討し、道民の意向を反映できる候補にしたい」と述べた。

この記事は北海道の知事選について述べた記事であるが、BCCWJ コーパスには元コーパスのタイムスタンプ情報が含まれていないため、この表現が指すイベントを同定するのは困難である。この事例においては、周辺文脈から「○一年度からの開発計画の進捗度合い」といった、記事が書かれた時間を推測できる手がかりを見つけることができたため、“2003 年北海道知事選挙” に関連付けたが、こういった特定のイベントに関するメンションをエンティティに紐付けるのは一般には難しい。同様の問題は「元大統領」といった地位／職業に関する言及や「ワールドカップ」等のスポーツイベントに関する言及でも見られた。

3 Wikification 実験

本研究で構築したデータセットを用いて、日本語 Wikification の評価実験を行う。Wikification を実現するためには、メンション m を検出するタスクと、m にリンクするエンティティ e を推定するタスクをどちらも解く必要があるが、本稿では正解のメンションが既知の状況のもとでエンティティを求める曖昧性解消タスクについて評価を行った結果を述べる。

曖昧性解消の手法として、アンカーテキストの確率分布に基づく手法 [10] を採用し、メンション m のエンティティ $\hat{e}$ を次式で推定する。

$$\hat{e} = \underset{e \in E}{\operatorname{argmax}}\, p(e|m) \tag{1}$$

ここで、E は日本語 Wikipedia に収録されている全エンティティの集合である。確率 $p(e|m)$ は、日本語 Wikipedia のアンカーテキストから、次式で求める。

$$p(e|m) = \frac{m \text{ が } e \text{ のアンカーとして出現する回数}}{m \text{ の総出現回数}} \tag{2}$$

この手法は、メンション m の出現文脈を無視して、日本語 Wikipedia コーパス内でよくリンクされているエンティティを選ぶことに相当する。非常に単純な手法であるが、Ling らは様々なデータセットで良好な成績を示すことを報告している [6]。なお、メンション m に対して $\forall e : p(e|m) = 0$ の場合は、NIL と判定する。

本研究で構築したコーパスに対して、式 1 による Wikification を適用したところ、正解率は 77.28%（22,514 件中 17,398 件正解）であった。したがって、文脈を全く考慮しない Wikification でも、比較的高い正解率が出ることが分かった。

曖昧性解消に失敗する要因をランダムにサンプリングした誤り事例から確認したところ、以下のような要因が大きな割合を占めることが分かった。

メンション・エンティティの意味クラスの相違

会場は 名古屋 City 正解:“名古屋市” 予測:“名古屋駅”・池下の愛知厚生年金会館。

今回用いている手法はメンションの意味クラス、エンティティの意味クラスを考慮していない。このパターンの誤りは多く、他にも例えば国名 / 言語名、人名 / 地名、地名 / スポーツチーム名などの誤りがみられた。メンションの意味クラスを詳細に予測することが EL の性能改善に寄与するという報告 [6] もあり、今後積極的に対応していきたい。

別名／リダイレクトの不足

夏にはジャスラック 正解:“日本音楽著作権協会” 予測:NIL の職員が店に説明に来た。

エンティティの別名や省略名が Wikipedia 内に十分に収録されていないことが原因となる事例も多かった。この例は、“日本音楽著作権協会” というエンティティの別名から「ジャスラック」という文字列が漏れているために起こる誤りである。

トピックの問題

声明によると、博士は核 正解:“核兵器” 予測:“細胞核” 拡散の責任を認めうえで、これまでの貢献を考慮して寛大な措置を取るよう求めたという。

この例のように、メンションが出現する文脈のトピックと、エンティティが属するトピックの整合性を確かめることで正しく曖昧性解消できる可能性がある事例も多くみられた。

Wikipedia 内のリンクの偏りに起因する問題

世界的に注目される日本人が増えており、野口 正解:“野口英世” 予測:“野口茂樹” (野球選手) の採用に「現在」をだぶらせ、親しみを深める向きもあろう。

今回用いた手法は、Wikipedia のアンカーテキストを用いて計算した確率値からエンティティを求めている。このため、スポーツ分野など、Wikipedia 内でリンクが密に張られているカテゴリに関連するエンティティがより優先されてしまうという問題がみられた。

4 まとめ

本論文では、日本語のエンティティリンキングの研究を進めるため、BCCWJ の新聞記事コーパス中に出現する固有表現に対して、日本語 Wikipedia へのリンクを付与したコーパスを構築した。データセットの構築やベースライン手法の実験結果から、日本語のエンティティリンキングが簡単なタスクでは無いことが明らかになった。本研究で構築したコーパスは、`http://www.cl.ecei.tohoku.ac.jp/jawikify` で配布されている。

本研究で実装したエンティティリンキング手法は、メンションの意味クラス、エンティティの意味クラス、メンションの文脈など、有用そうな手がかりを使っていない。Wikipedia の各記事に関根の拡張固有表現階層を割り当てる研究 [11] も進めており、この成果をエンティティリンキングに応用する予定である。

謝辞

この研究は、文部科学省受託研究「実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発」および文部科学省科研費（15H01702, 15H05318）の一環として行われた。

参考文献

[1] Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. Fast and space-efficient entity linking for queries. In *Proc. of WSDM*, pp. 179–188, 2015.

[2] Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. Entity disambiguation for knowledge base population. In *Proc. of COLING*, pp. 277–285, 2010.

[3] Hannaneh Hajishirzi, Leila Zilles, Daniel S. Weld, and Luke Zettlemoyer. Joint coreference resolution and named-entity linking with multi-pass sieves. In *Proc. of EMNLP*, pp. 289–299, 2013.

[4] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proc. of EMNLP*, pp. 782–792, 2011.

[5] Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. The impact of named entity normalization on information retrieval for question answering. In *Proc. of ECIR*, pp. 705–710, 2008.

[6] Xiao Ling, Sameer Singh, and Daniel Weld. Design challenges for entity linking. *TACL*, Vol. 3, pp. 315–328, 2015.

[7] Paul McNamee and Hoa Trang Dang. Overview of the TAC 2009 knowledge base population track. In *Text Analysis Conference (TAC)*, pp. 111–113, 2009.

[8] Rada Mihalcea and Andras Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proc. of CIKM*, pp. 233–242, 2007.

[9] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proc. of ACL-HLT*, pp. 1375–1384, 2011.

[10] Valentin I. Spitkovsky and Angel X. Chang. A cross-lingual dictionary for english wikipedia concepts. In *Proc. of LREC*, pp. 3168–3175, 2012.

[11] 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.

[12] 長田誠也, 末永圭吾, 善積正伍, 庄司和正, 吉田享晴, 橋本恭明. エンティティリンキングを用いたドキュメントに対する地点情報の付与とその応用. 言語処理学会第 21 回年次大会, pp. A4–4, 2015.

[13] 中村達哉, 白川真澄, 原隆浩, 西尾章治郎. ソーシャルメディアからの言語横断的な話題抽出に向けたエンティティリンキング手法. In *DEIM Forum 2015*, pp. A3–1, 2015.

[14] 古川竜也, 相良毅, 相澤彰子. 言語横断エンティティリンキングのための語義曖昧性解消. 情報知識学会誌, Vol. 24, No. 2, pp. 172–177, 2014.

[15] 林良彦, 山内健二, 永田昌明, 田中貴秋. 言語間の情報補完を用いた対訳文の Wikification. 2014 年度人工知能学会全国大会, pp. 1A2–2, 2014.