日本語エンティティリンキングコーパスの構築に向けた予備的アノテーション

東山翔平 * 内山将夫
国立研究開発法人情報通信研究機構 ユニバーサルコミュニケーション研究所
* 責任著者：shohei.higashiyama@nict.go.jp

概要

エンティティリンキングは，言語表現と，実世界の物や概念を表す知識ベース上のエントリとを対応付けるタスクである．同タスクのための言語資源の構築は，英語を中心に行われてきており，日本語のシステムの評価に利用できる言語資源は僅かである．本研究では，日本語エンティティリンキングシステムの評価に適したアノテーションコーパスの構築に向けて，設計方針とアノテーション基準を策定し，小規模なアノテーションを実施した．本稿では，それら方針・基準とともに，アノテーション作業プロセスと得られたデータの記述統計・特徴について報告し，今後の展望を述べる．

1 はじめに

エンティティと呼ばれる実世界の固有の物や概念に対し，人間は言語表現を介して言及する．**エンティティリンキング**（EL）では，そうした物や概念と文章中の言語表現を紐付ける問題を扱う．

EL は，**固有表現認識・メンション認識**と**エンティティ曖昧性解消**の二つの部分問題に分けられる．固有表現認識・メンション認識は，何らかのエンティティを指し示すテキスト中の言語表現を特定する処理であり，多くの場合，エンティティのカテゴリを同定する処理も含む．エンティティを指し示す言語表現は，（エンティティの）**メンション**と呼ばれ，特に固有名を中心とする言語表現の場合には**固有表現**とも呼ばれる．エンティティ曖昧性解消は，実世界の様々なエンティティに対応するエントリが収録された知識ベース，典型的には Wikipedia を前提とし，テキスト中のメンションを適切なエントリに対応付ける処理である．なお EL において，Wikipedia を知識ベースとして，認識したメンションを Wikipedia エントリ（記事）へ対応付ける問題設定は，特に Wikification [1] と呼ばれる．

EL の応用として，入力テキストのエンティティ曖昧性解消結果を考慮した文書分類や質問応答，曖昧性解消済みのエンティティ知識を用いた関係抽出や知識ベース拡充，獲得された知識や構築された知識ベースを利用した分散表現学習や言語生成などが挙げられる．

2000 年代後半以降，英語を中心に EL のための言語資源の構築 [2, 3, 4, 5, 6] や，そうした資源を用いたシステム評価の横断的研究 [7, 8, 9, 10] が行われてきた．一方，日本語 EL のために構築された言語資源は限られている [11, 12, 13].

最近の深層学習技術や基盤モデルの発展の中で，英語を中心とするデータから学習されたモデルが高い多言語処理能力を有する状況 [14] も見られ，一般的には言語個別のタスク特化システムを開発する必要性は低くなっている．しかしながら，日本語の EL タスクでは，日本特有の人物，組織，場所などの概念やそれらを表す言語表現を適切に解析し，相互に関連付けることが求められる．そこで本研究では，日本語 EL システムの正当な性能評価に必要となる，日本語の表現を豊富に含む日本語コーパスの構築について検討する．

EL 用コーパスの設計にあたって考慮すべき観点は，§3 で述べるようにいくつかある．その中でも，どのような場合にメンションと知識ベースエントリの対応関係を認定するかという点は自明ではなく，また従来研究でも必ずしも網羅的に議論されてきたわけではない．より具体的には，メンションが指す概念と直接的には合致せず，「全体と部分」や「前身と後身」のように，何らかの**概念的な重なり**のあるエントリが存在する場合に，対応関係の認定が難しい状況が生じる．本稿では，概念的重なりの種類を細分化した上で，その種類を表すタグとともにメ

Keywords: 自然言語処理，固有表現認識，エンティティ曖昧性解消，エンティティリンキング，コーパスアノテーション

Corpus	Public	Lang	Text Genre	Knowledge Base	Category	Nominal	Annotator
Ours	–	ja	Government site	Wikidata	Various	△	Human
jawikify [11]	✓	ja	News	Wikipedia	Various	✗	Human
JWC [12]	NY	ja	Blog, Twitter, etc.	Wikipedia	Various	✓	Human
SHINRA 2022 [13]	✓	ja	Wikipedia	Wikipedia	Various	✓	Auto
LRE Corpus [15]	△	ja	Twitter	Original	Geographic	✓	Human
ATD-MCL [16]	NY	ja	Travelogue	OpenStreetMap	Geographic	✓	Human
Mewsli-9 [17]	✓	Multi	Wikinews	Wikidata	Various	As Is	Article Editor
DaMuEL [18]	✓	Multi	Wikipedia	Wikidata	Various	As Is	Article Editor+Auto

表 1 日本語テキストを含む EL 用データセットの比較

ンションにエントリを割り当てるという一つのアノテーション基準を提示する．

設定したアノテーション基準に基づいて 12 記事 707 メンションへの予備的アノテーションを実施した結果，メンションと知識ベースエントリを直接的な合致として対応付けられるケースは約 60%にとどまった一方，概念的重なりを含めた合致として対応付けられるケースは約 80%となり，後者も含めてメンション・エントリ間の対応関係を認定する必要性が示唆された．

2 関連研究

日本語 EL に関する言語資源を構築した研究には，一般の EL に関する研究 [11, 12, 13] と，Geoparsing とも呼ばれる地理的な EL に関する研究 [15, 16] がある．また，日本語を含む多言語の EL のための言語資源構築の研究 [17, 18] も行われている．

Jargalsaikhan ら [11] による日本語 Wikification コーパス（jawikify）[1]，Murawaki ら [12] による Japanese Wikification Corpora（JWC）[2]は，ともに日本語 Wikification タスクのために構築された人手アノテーションコーパスである．Jargalsaikhan らは，拡張固有表現タグ付きコーパス [19] のうち BCCWJ [20] 新聞データを使用し，予め付与されているメンションに対して Wikipedia エントリの人手付与を行った．Murawaki らは，BCCWJ 白書・Yahoo!ブログデータ，Twitter データに対し，メンションと Wikipedia エントリの人手付与を行った．Murawaki らは，対応するエントリが存在する限り，あらゆるトピックのエンティティを対象とする方針としている．

Matsuda ら [15] による場所参照表現タグ付きコーパス（LRE コーパス）[3]，Higashiyama ら [16] による ATD-MCL[4]は，特定の場所を意図して書かれた言語表現（場所参照表現）と，それら地理的なメンションに対する地理データベース上のエントリを人手付与したコーパスである．両研究では，災害情報や観光情報の把握といった地理的な情報処理への応用を想定し，認識・曖昧性解消の対象とするメンションを地名・施設名等のカテゴリに限定している．

関根ら [13] は，Wikipedia 構造化のための森羅プロジェクト[5]において，日本語 Wikipedia 記事中の属性値を対象としたリンキングの共有タスク[6,7,8]を開催している．属性値のリンキングとは，拡張固有表現階層 [21] においてカテゴリ（例：「人名」）ごとに規定された属性（例：「作品」）を前提とし，Wikipedia 記事（例：カテゴリ「人名」の「島崎藤村」の記事）中の属性値メンション（例：カテゴリ「作品」の「嵐」）に対して適切な Wikipedia 記事（例：「嵐_(小説)」の記事）を選択する問題である．これらタスクでは，システムで自動アノテーションされたデータが評価に使用されている[9]．

Botha ら [17] は，多言語のテキスト集合に含まれる特定言語のメンションを，言語共通の知識ベース中のエントリに紐付ける Multilingual Entity Linking（MEL）タスクを提案し，MEL のための多言語データセット Mewsli-9[10]を構築した．Mewsli-9 は，日本語含む 9 言語の Wikinews 記事からなる．記事編集者により Wkipedia リンクが設定されたアンカーテキストをメンションとして，Wiki 間リンク

1 http://www.cl.ecei.tohoku.ac.jp/jawikify/
2 https://murawaki.org/research/wikify-data.html
3 http://www.cl.ecei.tohoku.ac.jp/~matsuda/LRE_corpus/
4 https://github.com/naist-nlp/atd-mcl
5 http://shinra-project.info/
6 http://shinra-project.info/shinra2021linkjp/
7 https://2022.shinra-project.info/data-download
8 https://2023.shinra-project.info/linking/
9 同プロジェクトは「協働による知識の構造化」を趣旨とし，システム出力の正しさを評価するというよりも，異なるシステム間の出力の一致や異なりを考慮することで，より高精度な知識ベースを構築することを意図していると考えられる．
10 https://github.com/google-research/google-research/tree/master/dense_representations_for_entity_retrieval/mel

（Wikimedia サイト間リンク）に基づいて Wikidata エントリ ID[11]が付与されている．

Kubeša ら [18] は，Wikipedia 記事間リンクと Wiki 間リンクに基づいて，53 言語の Wikipedia 記事に Wikidata エントリ（PER，ORG，LOC，EVENT，BRAND，WORK_OF_ART，MANUFACTURED の 7 カテゴリ）を付与したデータセット DaMuEL[12]を構築した．Wikipedia では，記事本文中で言及される各概念（エンティティ）について，最初に出現したメンションにのみ記事間リンクを設定する文書スタイルが採用されているが，Kubeša らは，2 回目以降の出現メンションに対しても，ヒューリスティックな基準を満たした場合にリンクを自動付与している．

以上の日本語または日本語を含む多言語の EL データセットと，本研究で構築しているコーパスの特徴を表 1 に示した[13]．データセットの特徴のうち，アノテーションの品質に直接的に影響するのが，アノテータの観点である．特に記事編集者が作成したメタデータを用いた場合，メンションとリンク先エントリの認定基準が記事編集者に委ねられる点，同一エンティティのメンションが基本的に一度しか付与されない点により，アノテーションの一貫性や網羅性が不十分となる[14]．

3 タスク定義に関わる論点

EL タスクは，素朴には「言語表現に対し，それが表す概念に “対応する” 知識ベースエントリを特定する問題」と定義される．しかし，どのような場合に言語表現とエントリの対応関係を認定すれば良いかという点は自明ではなく，メンションの認定基準と合わせて，エントリの認定基準が必要となる．

Ling [7] らは，EL の標準的なタスク定義が確立されていないこと指摘し，データの一貫性のためにアノテーションガイドラインで定義されるべき点として以下の 5 項目を挙げた．

1. **カテゴリ**：対象とするエンティティのカテゴリ．
2. **一般名詞句のメンション**：エントリ付与対象のメンションを固有名に限定するか，一般名詞句も含めるか．Ling らは，認定基準が主観に委ねられることから一般的な概念のメンションを含めるのは問題があると述べ，彼らの研究での評価対象を固有名のメンションに限定している．
3. **重なり合うメンション**：ネストするメンションを含め，スパンに重なりのあるメンションの扱いをどうするか．また，複合名詞句的なメンションを一つのメンションと扱うか複数のメンションと扱うか．
4. **メトニミー**：メトニミーのメンションに対し，表層と指示内容のどちらに基づいてエントリを割り当てるか．メトニミーは，「隣接関係に基づいて A でもって B に指示を横すべりさせる表現法」 [22] と定義される．たとえば，「WBC の決勝で日本が 3 度目の優勝を果たした．」という文での「日本」が，国家としての「日本 Q17」ではなく「野球日本代表 Q962145」という野球チームを指している例が該当する[15]．
5. **エントリの詳細度**：メンションにどのような詳細度のエントリを割り当てるか．たとえば，“United States Congress Q11268” と “116th United States Congress Q28227688” を区別するか．Ling らは，仮に “Joe Biden Q6279” と “United States Congress Q11268” の間に `President_Of` 関係を認定してしまうと，任意の議会期において Joe Biden が大統領であるという，事実に反する内容を表すことになり，関係抽出において問題になる点を指摘している．

項目 1〜3 については，コーパス構築時の目的やコストなどの制約によって様々な方針があり得る．項目 4 については，メトニミーかどうかの判断が難しい状況もあり得るものの，基本的には言語表現が指し示している概念に基づいてエントリを決定することが好ましいと考える．項目 5 は複合的な観点であり，具体的な状況へ細分化する余地があるが，関連する事例や議論が日本語 EL 研究の文献でも以下のように述べられている．

Jargalsaikhan ら [11] は，エントリの選択が問題になる例として，メンション “World Cup” に対するエン

11 Wikidata ではエントリにあたるコンテンツを「エンティティ」と呼んでいるが，本稿では，実世界の事物・概念そのものを「エンティティ」と呼び，知識ベース上の項目は「エントリ」と呼ぶ．

12 https://hdl.handle.net/11234/1-5047

13 データ公開状況（Public）について，“NY”（Not Yet）のデータは本稿執筆時点で未公開，“△” のデータは一部のみ公開されていることを表す．一般名詞句のメンション（Nominal）について，本研究では一部のみ対象とすることから “△” とし，Mewsli-9 および DaMuEL ではアンカーテキストをメンションとして採用していることから “As Is” と記載した．

14 網羅性が問題になるかは応用依存であり，たとえば文書単位での分類や検索では重要ではないと考えられる．

15 メンションまたはエントリについて，該当する Wikidata エントリがある場合には右上にその ID を付して表す．

トリ “FIFA World Cup” とエントリ “2002 FIFA World Cup” の選択や，メンション “Kobe Wing Stadium” に対する運営企業のエントリ “Kobe Wing Stadium Q11589735” と球技場の公式名称のエントリ “Misaki Park Stadium（御崎公園球技場）Q494766” の選択の例を挙げ，“Choose the entity that is the most specific in possible candidates.” という基準を設定している．

Murawaki ら [12] は，“Topical Matching”（話題の重なりに基づく一致の認定）を許容する基準を採用し，上位概念と下位概念，組織と組織の前身，「メーカー」と「製造業」のような類義語的な関係にあるようなメンション・エンティティ対が該当することを述べている．

森羅のリンキングタスク定義[16]では，メンション（属性値）とエントリ（Wikipedia 記事）との一致の度合いを「完全一致」と「準一致」に分け，準一致に該当するケースとして，(1) Later Name（後の名称），(2) Part-of（含まれている），(3) Derivation-of（形態違い）の三つを挙げている．(1) にはメンション「旭硝子」に対して同社の変更後の社名「AGC」のエントリを割り当てる例，(2) にはメンション「東京女子大学現代教養学部」に対して同組織の上位組織「東京女子大学」のエントリを割り当てる例，(3) には公演名のメンション「真夏の夜の夢」に対して文芸作品名の「夏の夜の夢」のエントリを割り当てる例が該当する．

エントリ選択の難しさを示している以上のような事例は，エンティティの概念的な重なりに該当するものと言える．つまり，社会総体的な概念の認定・名付けと，テキスト中での書き手個人の認識に基づく言語表現の使用の結果，「重なりのある概念が同一または重なりのある名称で言及される状況」が生じる．さらに，知識ベース編集者全体によって構築された知識ベースが不均一・不完全なものであることも，実際に選択可能なエントリに制限を課すことになる．

本研究では，これらの観点に対する方針を §4 で示しつつ，特に項目 5 について §5.2 で詳細な状況に分けて整理する．

4 本研究でのコーパス設計方針

本研究でエンティティとして想定するものは，実世界の人物・生物・抽象物・場所・出来事などのインスタンスであり，固有の名称を持ち，固有名などの言語表現によって指示対象となる個体や範囲が特定可能なものとする．この方針は，ドメイン非依存の情報抽出で対象とされるカテゴリを広く含む点と，割り当てるべきエントリを定めやすい点とのバランスを意識したものである．

知識ベースには，多言語 Wikipedia 記事等を情報源として構造化されたデータベースである Wikidata[17] [23] を採用する．

4.1 メンションの認定に関わる観点

前節で述べた Ling らの 5 項目のうち，項目 1〜4 はメンションの認定に関わるものである．これらに対する本研究の方針を以下のように定める．

項目 1：対象カテゴリ 拡張固有表現階層（Version 9）[18]の第 2 階層をベースとしたカテゴリを採用する．具体的には，拡張固有表現の「1 名前」に該当するカテゴリのうち，前述のエンティティに相当すると判断した，人名，生物呼称名，組織名，地名，施設名，プロダクト名の一部（§5.1.1 で後述），イベント名を対象とする．また，拡張固有表現の「2 時間表現」のうち，固有の名称が付けられた日付（例：「こどもの日」）や時代（例：「鎌倉時代」）にあたるものも対象とする．自然物名（化合物名・鉱物名のような物質名や，動植物名のような種名），病気名，色名と，プロダクト名の一部は，典型的なエンティティから外れると考え，現時点では対象外とした．

項目 2：対象メンションの形式 指示対象のエンティティが特定可能な固有名に加え，一般名詞句や指示代名詞の表現についても，文書中で固有名と共参照関係にある場合や，文脈から指示対象が特定できる場合には対象とする．固有名については，特定のインスタンスを指し示す呼称であると捉え，正式な名称のほか，非公式な名称・通称名，臨時的な呼称も含める．

項目 3：重なり合うメンションの扱い ネストの構造にある言語表現に対して最も外側の表現のみメンションとして扱う．ネスト構造への多重アノテーションは作業負荷が大きいためである[19]．

項目 4：メトニミーの扱い メトニミーは，エンティティ間の概念的な隣接性によって生じる状況であり，たとえば「日本」と「野球日本代表」との隣

16 https://drive.google.com/file/d/1PpaYRBvxrrqUPNuAC48l1h6fzIiyaxAM/view

17 https://www.wikidata.org/

18 http://ene-project.info/ene9/

19 将来的に，ネストの内側のメンションにもアノテーションを行う拡張も考えられる．

接によって，後者が前者の名称で言及されるということが起こる．これに対しては，言語表現の指示対象にあたるエンティティに基づいて，メンションとみなすかどうかと，メンションとみなす場合のカテゴリを判断する．

4.2 エントリの認定に関わる観点

Ling らによる項目 4 と 5 は，エントリの認定に関わる点である．これらについて以下の方針とする．

項目 4：メトニミーの扱い メンション認定の方針と同様に，メンションの指示内容に基づいてエントリを選択する方法を採用する．

項目 5：エントリの詳細度 項目 5 の中で Ling らが挙げた例は，エンティティ間の概念的重なりによって生じる状況の一種と言える．これに対し，概念的重なりのあるエントリ同士を区別しつつエントリ選択を行う方針とする[20]．

知識ベースの制約 また，所望のエントリが知識ベース上に存在するとは限らない点に対しては，(i) メンションの指示対象エンティティと合致するエントリが存在する場合のみ，エントリを選択する方法，(ii) そのようなエントリが存在しない場合に，概念的重なりのあるエントリの中で極力近いエントリを（補足情報付きで）選択する方法の二つが考えられる．本研究では (ii) を採用する．

以上をまとめると，メンションの指示対象として想定されるエンティティとほぼ合致する知識ベースエントリが存在する場合，そのエントリを割り当てることを優先し，想定されるエントリと概念的重なりがあるエントリが一つ以上存在する場合，「概念的重なり」の種類を表すタグとともに適当なエントリを割り当てる．これは，森羅リンキングタスクでの完全一致と準一致の区別と同様の方針であるものの，本研究ではより詳細な状況へ細分化している．

5 アノテーション基準

前述の設計方針（§4）の下で詳細化したアノテーション基準について述べる．なお，本基準は，予備的アノテーション作業（§6）と並行的に定めた試験的なものであり，今後のアノテーションデータの拡大に伴って変更が生じる可能性がある．

20 他にも，概念的重なりのあるエントリを同一視し，同一視したエントリのグループの中で，どのエントリを選択しても適切（EL システムにとって正解）とみなすような方法も考えられる．

Category	Our Tag	ENE Category
非固有名	NOM	–
人名	PER	1.1 人名
生物呼称名	LIV	1.3 生物呼称名
組織名	ORG	1.4 組織名
地名	LOC	1.5 地名
施設名	FAC	1.6 施設名
プロダクト名	PRO	1.7 プロダクト名（一部）
イベント名	EVE	1.8 イベント名
時間名	TIME	2.1 時間（固有名を持つもの）

表 2 カテゴリと拡張固有表現（ENE）との対応

5.1 メンションの認定

5.1.1 カテゴリの種類

本研究で採用したエンティティカテゴリを表 2 に示す．一般名詞句等のメンションには，それが指すエンティティのカテゴリによらず同一のタグ NOM を付与することとした．いずれのカテゴリについても，固有名を持つ事物・概念のみをエンティティとして想定しており，時間名に関しても数値等による時間表現はメンションとみなさない．プロダクト名については，拡張固有表現のプロダクト名のうち，固有名等によって指示対象となる範囲を特定しやすいと判断した以下を対象とし，その他は対象外とした．

1.7.1 株名，1.7.2 便名，1.7.3 識別番号，1.7.4 サービス名，1.7.5 ブランド名，1.7.6 ソフトウェア名，1.7.7 情報機器名，1.7.8 玩具名，1.7.9 楽器名，1.7.10 衣類名，1.7.11 医薬品名，1.7.12 キャラクター名，1.7.13 作品名，1.7.14 出版物名，1.7.15 ゲーム名，1.7.16.0 食べ物名_その他，1.7.17 武器名，1.7.18 乗り物名，1.7.19.3 賞名，1.7.19.4 勲章名，1.7.19.6 技術名，1.7.20.1 条約名，1.7.20.2 法令名，1.7.24 バーチャルアドレス名

5.1.2 スパンの認定

メンションのスパンの決定について，本研究で採用した基本的な基準は以下の通りである．例文中では，メンションとみなさない表現を波線で示す．

1. **メンションのネスト**：メンションがネスト（入れ子）構造になった複合名詞句等の表現については，対象カテゴリに相当する表現の中で，最も外側の表現を一つのメンションとし，それより内部の表現にはアノテーションを行わない．
   (a) 三陸・海の博覧会 EVE
   (b) 外務省 ORG 軍縮不拡散・科学部 ORG 長
2. **メンションの分割**：全体として一つのエンティ

ティを指す正式名称等と考えられる場合には一つのメンションとする．そうでなければ複数のメンションに分ける．住所等で地名が連続する場合には，行政区またはそれより大きい地域区分は別のメンションに分け，それより小さい地域名・番地等は一つのメンションとする．

(c) 京都市動物園 FAC

(d) 伊藤園 ORG 健康ミネラルむぎ茶 PRO

(e) 山梨県 LOC 北杜市 LOC 高根町長澤 2141 LOC

3. **メンションの属性的情報**：同格構造やコピュラ文において，メンションの他にその属性的な情報を表す表現が出現する場合，後者の表現はメンションとしない．ただし，同格的に表れる修飾句も含めて正式名称等と捉えられる場合，全体を一つのメンションとする．

(f) 日本最北端の駅「稚内駅 FAC」

(g)「南町田グランベリーパーク FAC」は，(中略) 約 20ha のエリアで，（例文 MLIT-1-1:5）

(h) 道の駅 針 T・R・S（はりてらす）FAC

5.2 エントリの認定

前述のように，メンションの指示対象として想定されるエンティティと概念的重なりがあるようなエントリのみ存在する場合，「概念的重なり」の種類を表すタグ（CO タグ）とともに該当するエントリを割り当てる方針とした．概念的重なりの代表的な状況として，(1) 全体と部分（CONTAINS タグまたは PART_OF タグ），(2) 周期的なシリーズとインスタンス（PERIODIC_SERIES_OF または PERIODIC_INSTANCE_OF タグ），(3) エンティティの時間的変化（DIACHRONIC_CORRESPONDENCE タグ），(4) 概念的隣接・複合および (5) その他の概念的重なり（OTHER_CONCEPTUAL_OVERLAP タグ），(6) 概念定義の曖昧さ（VAGUE タグ），(7) 複数の種類の概念的重なり（MULTI_CONCEPTUAL_OVERLAP タグ）の 7 通りに整理した．このそれぞれについて，以下で例とともにアノテーション基準を説明する．

例では，該当する Wikidata エントリが存在するメンションを「国土地理院 Q2986578」のように Wikidata の ID を付して表し，Wikidata エントリが存在しないメンションを「日本三大松原 NIL」のように表す．実際の文を例文として挙げる場合には付録 A.1 中の文 ID を併記した．

5.2.1 全体と部分

二つのエンティティが全体と部分の関係にある状況では，全体エンティティまたは部分エンティティのエントリが存在しない場合に問題となる．

ケース 1：例 (i) の「四国地方測量部」のように部分エンティティを指すメンションに対し，合致する部分エンティティのエントリがなく，全体エンティティのエントリがある場合には，PART_OF タグとともに全体エンティティのエントリを付与する．

ケース 2：例 (k) の「日本三大松原」のように全体エンティティを指すメンションに対し，合致する全体エンティティのエントリがなく，部分エンティティのエントリがある場合には，CONTAINS タグとともにいずれかの部分エンティティのエントリを付与する．

(i)「国土地理院 Q2986578」（例文 GIS-1-1:9）が全体，「四国地方測量部 NIL」（例文 GIS-1-1:10）が部分に相当する．

(j)「中国・四国地方 Q5358181」が全体，「中国地方 Q127864」「四国 Q13991」，がそれぞれ部分に相当する．

(k)「日本三大松原 NIL」が全体，「三保の松原 Q869275」，「気比松原 Q11548084」，「虹の松原 Q5370854」がそれぞれ部分に相当する．

5.2.2 周期的なシリーズとインスタンス

周期的に開催される催し物やそれに伴って編成される組織などは，特定の日時・期間に開催・編成される個々のインスタンスのエンティティと，個々のインスタンスを包括するシリーズ全体についてのエンティティが想定できる．

ケース 1：例 (l) のようにシリーズ・インスタンス両方のエントリが存在する場合など，メンションが指すエンティティに合致するエントリがある場合はそれを付与する．

ケース 2：例 (m) の「つくば科学フェスティバル」のように，メンションが指すインスタンスのエントリが存在せず，シリーズのエントリが存在する場合には，PERIODIC_INSTANCE_OF タグとともにシリーズのエントリを付与する．

ケース 3：メンションが指すシリーズのエントリが存在せず，インスタンスのエントリが存在する場合は，PERIODIC_SERIES_OF タグとともにインスタンスのエントリを付与する．

(l)「夏季オリンピック Q159821」は 4 年おきに開催されるイベントのシリーズに相当し，「2020 年東京オリンピック Q181278」は 2021 年開催（2020 年から 1 年延期）のイベントのインスタンスを指し示す．

(m)「同フェスティバル」（＝つくば科学フェスティバル Q11272282）（例文 GIS-1-2:2）は毎年開催されるイベントのシリーズに相当し，「つくば科学フェスティバル NIL」（例文 GIS-1-2:1）は 2022 年開催のイベントのインスタンスに相当する．

5.2.3 エンティティの時間的変化（前身と後身）

エンティティの名称や実態が時間的に変化し，前身と後身の関係にあたるエンティティが生じる状況がある．たとえば，組織やサービスの名称変更や，市町村の統廃合による歴史的変遷などである．

ケース 1：例 (n) の「東京電力」や例 (o) の「三戸城」，例 (p) の「釜石市」などのように，メンションが指すエンティティに合致するエントリが存在する場合は，それを付与する．

ケース 2：仮に，例 (n) に関してかつての「東京電力株式会社」を指すメンションや，例 (p) に関して (γ)「旧釜石市」を指すメンションがあった場合，それらメンションの指すエンティティと直接的に合致するエントリはないものの，存在するエントリ（「東京電力ホールディングス Q333894」，「釜石市 Q329790」）が表すエンティティとの間で，エンティティとしての本質的内容・性質は維持されていると捉えることができそうである．この場合には，通時的な対応関係を表す DIACHRONIC_CORRESPONDENCE タグとともに該当するエントリを付与する．

ケース 3：例 (o) のメンション「城山公園」が指すエンティティに合致するエントリはなく，それと関連のあるエントリとしては「三戸城 Q10866038」がある．両者が表すエンティティは，地理的範囲や役割の違いが大きく，エンティティとしての本質的内容・性質が異なると捉えられそうである．この場合には，エントリの付与を行わない．

ケース 4：メンションが市町村よりも細かい地域名に相当する場合，歴史的地名も絡み，割り当てるエントリとして複数の候補が存在し得る．基本的には，メンションが現存する地名，歴史的地名のいずれを指す場合も，直接的に対応するエントリが存在する場合はそれを割り当てる．例 (h) において (α)，(β)，(γ) を指すメンションがあった場合，「釜石市 Q329790」を DIACHRONIC_CORRESPONDENCE タグとともに付与する．例 (q) のメンション「日頃市町」は，現存する地名かつ市町村よりも細かい地域を指す例であり，それを包含する市町村のエントリ（「大船渡市 Q384934」）を PART_OF タグとともに付与するか，またはその通時的な対応物を表すエントリ「日頃市村 Q11510764」を DIACHRONIC_CORRESPONDENCE タグとともに付与するという選択肢があり，いずれも可とする．

(n)「東京電力 NIL」は，2016 年 4 月の社名変更・体制変更後「東京電力ホールディングス Q333894」となっている．「東京電力」（例文 METI-1:2）は 2022 年に書かれた文章中でのメンションであり，後者を指すと捉えられる．

(o)「三戸城 Q10866038」は近世に存在した城，「国史跡三戸城跡 城山公園 NIL」は跡地周辺を領域とする公園である．「三戸城」と「三戸古城」（例文 MLIT-4-2:4）は存在していた当時の城，「城山公園」（例文 MLIT-4-2:5）は現代の公園を指すメンションと捉えられる．

(p) 歴史上「釜石」と呼ばれた地域として，町村制施行前に存在した (α)「釜石村 NIL」，1889 年に町村制の下で成立した (β)「釜石町 NIL」，1937 年に市制移行により成立した (γ)「釜石市 NIL」（「旧釜石市」），1955 年に市町村合併により現在の市域に拡大した (δ)「釜石市 Q329790」がある[21]．「釜石村」（例文 TLDB-1:74）と「釜石」（例文 MLIT-TLDB-1:60）などは (α) を指すメンション，「釜石」（例文 TLDB-1:71）と「釜石市」（例文 TLDB-1:74）は (δ) を指すメンションと捉えられる．

(q) メンション「日頃市町 NIL」（例文 TLDB-1:9）は，「大船渡市 Q384934」にあり，岩手県気仙郡にあった「日頃市村 Q11510764」の地域に相当する．

5.2.4 概念的隣接・複合

メトニミーのように概念的隣接に基づいてメンションと指示対象にずれが生じている状況や，「日本」が国家，統治機構，地理的領域といった複数の側面・属性を併せ持つというように，同一の（または重なりのある）名称で複合的な概念が指示される状況がある．この隣接と複合は，必ずしも明確に区別できない場合もあり得る．

21 http://mujina.sakura.ne.jp/history/03/index2.html

ケース 1：メンション「日本」が「野球日本代表 Q962145」を指すようなメトニミーの状況で，指示対象に合致するエントリが存在する場合，そのエントリを付与する．

ケース 2：例 (r) については，自治体組織としての函館市と地名としての函館市を表すエントリが単一のエントリとして存在する．このような場合，特定の側面に着目したメンションであっても，そのエントリを付与するので良い（メンション「市」にエントリ「函館市 Q26418」を付与）．

ケース 3：例 (s) や例 (t) については，国家としての日本国，地理的領域としての日本列島，統治機構としての日本政府を表すエントリがそれぞれ分かれて存在する．このような場合，メンションの指示対象に最も合致するエントリを付与する（たとえば例 (s) のメンション「日本」にはエントリ「日本列島 Q841337」を付与）．

ケース 4：メンションが指すエンティティと直接的に合致するエントリがなく，概念的重なりのあるエントリしか存在しない場合，OTHER_CONCEPTUAL_OVERLAP タグとともに該当エントリを付与する．たとえば，例 (u) で 2021 年上演のミュージカル作品「魔女の宅急便」を指すメンションがあった場合，同タグとともに原作を表すエントリ「魔女の宅急便 Q1768944」を付与する[22]．

(r)「市」（＝函館市）（例文 MLIT-2:56）は，自治体組織としての側面に着目したメンションである．「函館市 Q26418」含め，市町村を表すエントリは，市町村の組織としての側面と地名としての側面の両方を併せ持ち，エントリとして分離されていない．

(s)「日本 Q17」は日本の国家を指すエントリ，「日本列島 Q841337」は日本の領土に概ね相当する領域の地形を指すエントリである．「日本」（例文 GIS-1-2:5）は日本の地形，つまり日本列島を指すメンションと捉えられる．

(t)「日本政府 Q1190574」は日本の政府を指すエントリである．「日本」（例文 METI-1:2）は日本政府を指すメンションと捉えられる．

(u) メディアミックス作品「魔女の宅急便」については，角野栄子による児童文学作品「魔女の宅急便 Q1768944」，宮崎駿による 1989 年のアニメ映画作品「魔女の宅急便 Q196602」，清水崇による 2014 年の実写映画作品「魔女の宅急便 Q15260709」，その他のミュージカル作品「魔女の宅急便 NIL」がある．

5.2.5 その他の概念的重なり

例 (v)，(w) は，実際にはアノテーションに迷う事例ではないものの，上述した状況に当てはまらないような概念的重なりの事例があり得ることを示唆している．メンションに直接対応するエンティティのエントリが存在せず，一定の概念的重なりがあるエントリが存在する場合，OTHER_CONCEPTUAL_OVERLAP タグとともに該当するエントリを付与する．

(v)「東日本大震災 Q1136168」は，「東北地方太平洋沖地震 Q36204」による災害及びこれに伴う原子力発電所事故による災害を指す．メンション「東日本大震災」（例文 GIS-1-1:6）は文字通り前者を指すと捉えられる．

(w)「ニッポンフードシフト NIL」は農林水産省が実施する国民運動であり，開催の場所や期間が定まる具体的なイベントではない．同運動の一環として「食から日本を考える．NIPPON FOOD SHIFT FES. 山梨 NIL」という催しが開催されている．例文 MAFF-1:2 と MAFF-1:3 では，文字通りに前者と後者を指す表現がそれぞれ出現する．

5.2.6 概念定義の曖昧さ

書き手が言及している概念が曖昧なものであったり，公式な定義のうちどれに該当するのか不明確であったりする状況がある．北本は，この状況を Vagueness と呼び，同一の表現が複数の指示対象を指し得る Ambiguity の状況と区別している[23]．例 (y)，(z) のように，メンションに対応する概念定義が想定でき，またそれに対応するエントリが存在する場合は，そのエントリを付与する．例 (x) のように，漠然とした概念を指すメンションと捉えられる場合は，VAGUE タグとともに近いエントリを付与する．

(x)「尾道」に関するエントリには，「尾道市 Q696694」の他，旧尾道市市街地の歴史等についての「尾道 Q24887789」がある．「私は今尾道にいる，という安心と幸せ」という感覚を述べた文におけるメンション「尾道」（例文 MLIT-1-2:14）は，書

22 このエントリに対応する Wikipedia 記事はミュージカル作品についての記述を含む．

23 http://agora.ex.nii.ac.jp/~kitamoto/research/publications/geonlp23.html.ja

き手以外にとってどの概念を指しているのか明らかでない．

(y)「京都」に関するエントリには，「京都府 Q120730」や「京都市 Q34600」の他，歴史的な都市や地名としての「京都 Q740246」がある．歴史的な事実を述べた文におけるメンション「京都」（例文 MLIT-4-2:20）は，三つ目の「京都 Q740246」が最も合致するという解釈があり得る．

(z) 山としての「八ヶ岳」の定義は複数あり，長野県と山梨県の地域にまたがる「八ヶ岳連峰 Q11390433」全体を指す場合や，八ヶ岳連峰のうち山梨県を中心とする夏沢峠以南の範囲を指して「（南）八ヶ岳 Q905588」と呼ぶ場合がある．メンション「八ヶ岳」（例文 MAFF-1:12）が何を指すかは明らかでないが，山梨県の地域やイベントについて述べている文脈を考慮すると後者を指している可能性が考えられる．

5.2.7 複数の種類の概念的重なり

上述の概念的重なりのうち複数の種類に該当する場合，MULTI_CONCEPTUAL_OVERLAP タグとともにエントリを付与する．

6 予備的アノテーション

アノテーション基準の定義と並行して，第一著者により小規模なアノテーション作業を行った．本節では，アノテーション作業の過程と，得られたデータの記述統計を報告し，アノテーション揺れが生じやすいと想定される点について議論する．

6.1 アノテーションプロセス

記事選択　アノテーション対象記事として，日本の府省庁およびその内部部局のサイト[24]上で公開されている HMTL または PDF 文書を選択した．これらのサイトを選んだ理由は，(1) 出典の記載などの条件の下，翻案・公衆送信含む自由度の高い利用が認められている点[25]，(2) 日本国内における地域，出来事，施策，歴史などを説明・報告した文書が多く存在し，日本特有のエンティティを指し示すメンションの事例が豊富に得られる点による．具体的には，様々なエンティティのメンションを含むことなどを理由に，著者の主観により 10 文書 12 記事（付録 A.1 に記載）を選択した．

24　https://www.gov-online.go.jp/topics/link/index.html
25　「政府広報オンライン」上のコンテンツに関する利用条件：https://www.gov-online.go.jp/etc/tos.html

	Doc	Sent	Men	Cls
Total	12	362	707	413
Per doc	–	30.2	58.9	34.4

表 3　アノテーションを行った 12 記事の記述統計（Doc：記事数，Sent：文数，Men：メンション数，Cls：共参照クラスタ数）

テキスト整形・前処理　選択した原記事ファイルから，本文に該当するテキストを抽出し，手作業で文単位で整形した．各記事テキストに GiNZA[26] [24]（ja_ginza）を適用して固有表現抽出を行い，拡張固有表現に基づく固有表現タグを本研究で用いるタグに変換した後，さらにアノテーションツール brat [25] の入力ファイル形式に変換した．

メンションアノテーション　brat を用いて，各記事中のメンションを認定し，カテゴリタグを付与する作業を行った．

リンクアノテーション　各記事中のメンションについて，同一のエンティティを指すメンション全体（共参照クラスタ）に対して同一の Wikidata エントリ（ない場合には Wikipedia エントリ）の URL を，概念的重なりの種類を表す CO タグとともに割り当てる作業を行った．作業方法として，(1) brat 上で，共参照クラスタに含まれるべきメンション対に共参照関係を表す 2 項関係エッジを付与しつつ，共参照クラスタ中のいずれかのメンションに URL と CO タグを付与する方法[27]と，(2) 所定の形式の TSV ファイル上で，共参照クラスタに含まれるべき一連のメンションを連続する行として並べる方法のどちらでも良いとし，実際には両者を組み合わせた方法を用いた[28]．エントリの検索には，ウェブ検索ならびに Wikidata および Wikipedia サイト上の検索機能を用いた．

6.2 アノテーションデータの記述統計

全般的な記述統計　12 記事合計 362 文へアノテーションを行った結果，表 3 に示すように，メンション延べ数 707（異なり数 431），共参照クラスタ数

26　https://github.com/megagonlabs/ginza
27　brat 作業画面では，Notes 欄に URL を入力し，メンションの Entity Attribute として CO タグを選択した．
28　brat のみでも作業可能だが，同一記事中に大量のエッジを付与すると画面表示が複雑になり，共参照クラスタ内の全メンションを視覚的に把握するのが容易でなくなるため，brat での作業後に，最終的に TSV ファイルとして整形する方法を採用した．

	Total	WD	WD+	WP	WP+	NIL
Men	707	408	157	2	1	139
Cls	413	212	96	2	1	114

表 4 メンション（Men）および共参照クラスタ（Cls）へ付与されたエントリの状況別の件数．Wikidata/Wikipedia エントリを CO タグなしで付与（WD/WP），Wikidata/Wikipedia エントリを CO タグとともに付与（WD+/WP+），いずれも付与なし（NIL）の件数を表示．

	PART	P-INS	DIAC	OTHER	VAGUE	MULTI
Men	55	2	68	17	10	5
Cls	38	1	35	15	4	3

表 5 Wikidata エントリと CO タグの両方を付与された事例における CO タグの分布

413 となった．メンションおよび共参照クラスタに付与されたエントリの状況別の件数を表 4 に示す．全メンションのうち，Wikidata エントリを CO タグなし，CO タグありで付与されたものの割合はそれぞれ約 58%（408/707），80.0%（565/710）であり，直接的に対応するエントリが存在するメンションの割合は 6 割程度に限られる結果となった．したがって，直接的な対応に加えて概念的重なりのあるエントリも特定するタスク設定とする方が，EL システムの応用上の有用性は高くなると考えられる．全共参照クラスタのうち，Wikidata エントリを CO タグなし，CO タグありで付与されたものの割合はそれぞれ約 51%（212/413），75%（308/413）であり，メンションの場合と似た分布であった．Wikidata エントリがなく代替的に Wikipedia エントリが付与されたメンションおよび共参照クラスタはともに 3 件あった（うち 1 件は PART_OF タグも付与）．

CO タグ付与の内訳 Wikidata エントリとともに CO タグが付与された事例（表 4 の WD+）における CO タグの分布を表 5 に示す．メンション，共参照クラスタ単位ともに，PART_OF（PART）と DIACHRONIC_CORRESPONDENCE（DIAC）の二つが全体の 80%弱を占めた．両タグが付与されたメンションのカテゴリはともに LOC が最多であり，市町村よりも細かい粒度の地名や，歴史に言及する文脈で出現した地名に対して付与された事例が多かった．二つの CO タグに次いで，OTHER_CONCEPTUAL_OVERLAP（OTHER）と VAGUE が多く，PERIODIC_INSTANCE_OF（P-INS）と MULTI_CONCEPTUAL_OVERLAP（MULTI）は少数存在し，PERIODIC_SERIES_OF と CONTAINS が付与された事例はなかった．

Category	#	WD	WD+	Both	Example
NOM	53	60.4	18.9	79.3	本施設
PER	29	62.1	0	62.1	大島高任
LIV	2	0	0	0	関根の松
ORG	84	53.6	19.1	72.6	経済産業省
LOC	339	67.6	27.7	95.3	三戸
FAC	119	45.4	23.5	68.9	つくばカピオ
PRO	41	19.5	17.1	36.6	地理院地図
EVE	20	45.0	10.0	55.0	東日本大震災
TIME	20	65.0	0	0	世界津波の日
Overall	707	57.7	22.2	79.9	

表 6 カテゴリごとのメンション数（#），リンク率（WD，WD+，Both）と，メンションの実例．

カテゴリ別メンション数・リンク率 表 6 に，カテゴリ別のメンション件数と，Wikidata エントリ（WD）または Wikipedia エントリと CO タグ（WD+）を付与されたメンションの割合（リンク率）を示す．メンション数は地名（LOC），施設名（FAC），組織名（ORG）の順に多く，これら 3 カテゴリの計 542 メンションで全メンションの約 77%を占めた．カテゴリごとのメンションのリンク率は，プロダクト名（PRO）と，件数が少ない生物呼称名（LIV）を除き，CO タグなしの場合（WD）に 45〜68%程度，CO タグありも含めた場合（Both）に 55〜95%程度となった．プロダクト名のリンク率が低い点については，「邦内郷村志」（地誌名）や「宮古浦善宝丸」（舟名）といった歴史的なエンティティに対するエントリがなかったことや，「八ヶ岳地産地消カレー」や「放射線環境影響評価報告書」といった広く流通しているプロダクトではない臨時的・局所的なエンティティの言及もメンションとみなしてアノテーションしたことが主な理由として挙げられる．

6.3 議論

今後，複数名でのアノテーションの実施と作業者間一致率の評価を実施する予定である．現在のアノテーション基準に基づく作業でアノテーション揺れが生じやすいと想定される点として，複数のエントリが候補となる状況でのエントリ選択が挙げられる．たとえば，§5.2 例 (q) の「日頃市町」では所在地である「大船渡市 Q384934」(+PART_OF) と村制施行時の「日頃市村 Q11510764」(+DIACHRONIC_CORRESPONDENCE) が候補となる．今回のアノテーションでは，作業者

の主観により一つを選択する手軽な作業方法を採用した．しかし，データの一貫性や，システム評価用データとして用いる観点から考えると，同程度の妥当さのエントリは漏れなく付与されていることが望ましい．該当する複数のエントリを選択する作業方法も考えられるものの，作業の複雑さを抑えるための工夫も必要となる．

7 おわりに

本稿では，日本語エンティティリンキング（EL）コーパスの構築に向けて，コーパス設計方針とアノテーション基準の定義に基づき，予備的アノテーションを実施した結果を報告した．今後，アノテーションデータの規模を拡大した上で公開し，同データを用いた日本語 EL システムの評価・分析に取り組む予定である．

参考文献

[1] Rada Mihalcea and Andras Csomai. Wikify! Linking documents to encyclopedic knowledge. In Proceedings of the 16th ACM Conference on Information and Knowledge Management, pp. 233–242, 2007.

[2] Silviu Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 708–716, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[3] Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie M. Strassel. An evaluation of technologies for knowledge base population. In Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta, Malta, May 2010. European Language Resources Association.

[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 Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 782–792, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[5] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to Wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 1375–1384, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[6] Giuseppe Rizzo, Marieke van Erp, Julien Plu, and Raphaël Troncy. Making sense of microposts (#microposts2016) named entity recognition and linking (NEEL) challenge. In Proceedings of the 6th Workshop on 'Making Sense of Microposts' co-located with the 25th International World Wide Web Conference.

[7] Xiao Ling, Sameer Singh, and Daniel S. Weld. Design challenges for entity linking. Transactions of the Association for Computational Linguistics, Vol. 3, pp. 315–328, 2015.

[8] Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In Proceedings of the 10th International Conference on Language Resources and Evaluation, pp. 4373–4379, Portorož, Slovenia, May 2016. European Language Resources Association.

[9] Michael Röder, Ricardo Usbeck, and Axel-Cyrille Ngonga Ngomo. GERBIL – Benchmarking named entity recognition and linking consistently. Semantic Web, Vol. 9, No. 5, p. 605–625, January 2018.

[10] Marcel Milich and Alan Akbik. ZELDA: A comprehensive benchmark for supervised entity disambiguation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2061–2072, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[11] Davaajav Jargalsaikhan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. Building a corpus for Japanese wikification with fine-grained entity classes. In Proceedings of the ACL 2016 Student Research Workshop, pp. 138–144, Berlin, Germany, August 2016. Association for Computational Linguistics.

[12] Yugo Murawaki and Shinsuke Mori. Wikification for scriptio continua. In Proceedings of the 10th International Conference on Language Resources and Evaluation, pp. 1346–1351, Portorož, Slovenia, May 2016. European Language Resources Association.

[13] 関根聡, 中山功太, 隅田飛鳥, 渋木英潔, 門脇一真, 三浦明波, 宇佐美佑, 安藤まや. 森羅タスクと森羅公開データ. 言語処理学会 第 29 回年次大会 発表論文集, 2023.

[14] OpenAI. GPT-4 technical report. arXiv:2303.08774, 2023.

[15] Koji Matsuda, Akira Sasaki, Naoaki Okazaki, and Kentaro Inui. Geographical entity annotated corpus of Japanese microblogs. Journal of Information Processing, Vol. 25, pp. 121–130, 2017.

[16] Shohei Higashiyama, Hiroki Ouchi, Hiroki Teranishi, Hiroyuki Otomo, Yusuke Ide, Aitaro Yamamoto, Hiroyuki Shindo, Yuki Matsuda, Shoko Wakamiya, Naoya Inoue, Ikuya Yamada, and Taro Watanabe. Arukikata travelogue dataset with geographic entity mention, coreference, and link annotation. arXiv:2305.13844, 2023.

[17] Jan A. Botha, Zifei Shan, and Daniel Gillick. Entity linking in 100 languages. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 7833–7845, Online, November 2020. Association for Computational Linguistics.

[18] David Kubeša and Milan Straka. DaMuEL: A large multilingual dataset for entity linking. arXiv:2306.09288, 2023.

[19] 橋本泰一, 中村俊一. 拡張固有表現タグ付きコーパスの構築-白書, 書籍, Yahoo! 知恵袋コアデータ-. 言語

処理学会 第 16 回年次大会 発表論文集, 2010.

[20] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, Vol. 48, No. 2, pp. 345–371, 2014.

[21] Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. Extended named entity hierarchy. In Proceedings of the 3rd International Conference on Language Resources and Evaluation, Las Palmas, Canary Islands - Spain, May 2002. European Language Resources Association.

[22] 瀬戸賢一, 宮畑一範, 小倉雅明. ［例解］現代レトリック事典. 大修館書店, 2022.

[23] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. Communications of the ACM, Vol. 57, No. 10, p. 78–85, September 2014.

[24] 松田寛, 大村舞, 浅原正幸. 短単位品詞の用法曖昧性解決と依存関係ラベリングの同時学習. 言語処理学会 第 25 回年次大会 発表論文集, 2019.

[25] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 102–107, Avignon, France, April 2012. Association for Computational Linguistics.

A　付録

A.1　アノテーションに用いた記事と文の例

アノテーションに用いた 12 記事の情報と，記事内の文の一部を表 7 および表 8 に示す．“GIS-1-1” などの記事 ID の後ろの “T” は記事タイトル，“U” は記事 URL，その他の数値は本文であることを表している．本文テキスト中で，§5 で例として挙げたメンションを色付きで表示し，人名にあたる個所を伏字「■」にしている（これらは著者による編集である）．

GIS-1-1:T	「令和 4 年度大規模津波防災総合訓練」に参加
GIS-1-1:U	https://www.gsi.go.jp/common/000245996.pdf
GIS-1-1:1	11 月 13 日に高知県で南海トラフ地震を想定した大規模津波防災総合訓練が、国土交通省をはじめとする防災関係機関など 107 団体が参加して大規模に行われました。
GIS-1-1:6	パネル展示では、国土地理院の防災への取り組みや東日本大震災での対応、自然災害伝承碑の取り組みを紹介し、訓練会場付近のデジタル標高地形図や南海トラフの余色立体図では 3D メガネで立体的に高さの違いを体感していただきました。
GIS-1-1:9	国土地理院では今後も防災関係機関との連携強化、災害対応能力の向上及び防災教育・啓発を図るため、本訓練に積極的に参加していきます。
GIS-1-1:10	（四国地方測量部）
GIS-1-2:T	3 年ぶりの「つくば科学フェスティバル 2022」に出展
GIS-1-2:U	https://www.gsi.go.jp/common/000245996.pdf
GIS-1-2:1	11 月 12 日、つくばカピオ（茨城県つくば市）において開催された「つくばサイエンスコラボ 2022 科学と環境のフェスティバル」の「つくば科学フェスティバル」に出展しました。
GIS-1-2:2	同フェスティバルは毎年行われていましたが、新型コロナウイルス感染症の影響により 3 年ぶりの開催となりました。
GIS-1-2:5	国土地理院は、「日本の凹凸を体験しよう！」と題したブースを出展しました。
MAFF-1:T	「食から日本を考える。NIPPON FOOD SHIFT FES. 山梨」を開催！
MAFF-1:U	https://www.maff.go.jp/j/press/kanbo/anpo/230113.html
MAFF-1:2	農林水産省では、食と農のつながりの深化に着目した国民運動「食から日本を考える。ニッポンフードシフト」を実施しています。
MAFF-1:3	この取組の一環として、令和 5 年 1 月 22 日（日曜日）、山梨県 八ヶ岳南麓の旧北杜市立高根北小学校において、消費者と生産者、食品事業者が出会い、交わることで、日本の「食」が抱える課題や目指す未来について、ともに考えるきっかけとするイベント「食から日本を考える。NIPPON FOOD SHIFT FES. 山梨」を開催します。
MAFF-1:10	開催場所：旧北杜市立高根北小学校（山梨県 北杜市 高根町長澤 2141）
MAFF-1:12	八ヶ岳南麓の山梨県北杜市は、大消費地から近い立地条件を持ち、日本一長い日照時間、清らかで豊富な水資源など、豊かな自然環境を有しています。
METI-1:T	東京電力福島第一原子力発電所の ALPS 処理水の現状に関する韓国政府向けのテレビ会議説明会を開催しました
METI-1:U	https://www.meti.go.jp/press/2022/12/20221222003/20221222003.html
METI-1:2	日本側から、■■■■経済産業省資源エネルギー庁原子力事故災害対処審議官、■■■外務省軍縮不拡散・科学部長のほか、原子力規制庁、環境省、水産庁、東京電力が、また、韓国側から、■■■■■■■外交部気候環境科学外交局長のほか、国務調整室、科学技術情報通信部、海洋水産部、原子力安全委員会等の関係省庁が参加しました。
MLIT-1-1:T	大賞　国土交通大臣賞　南町田グランベリーパーク地区
MLIT-1-1:U	https://www.mlit.go.jp/common/001347545.pdf
MLIT-1-1:5	「南町田グランベリーパーク」は、東京都町田市の端にある東急田園都市線「南町田グランベリーパーク駅」（2019 年 10 月 1 日に「南町田駅」から改称）南側の約 20ha のエリアで、地元自治体と鉄道事業者の強力なパートナーシップのもと、都市基盤・商業施設・都市公園などを一体的に再整備・再構築し「新しい暮らしの拠点」の創出に取り組んできた。

表 7　アノテーションに用いた記事の情報（1〜5 記事目）

MLIT-1-2:T	優秀賞　「都市景観の日」実行委員会 会長賞　尾道市景観地区（尾道・向島歴史的風致地区）
MLIT-1-2:U	https://www.mlit.go.jp/common/001347545.pdf
MLIT-1-2:14	この感覚が、私は今尾道にいる、という安心と幸せとなる。
MLIT-2:T	第２章 地域資源を活用した良好な景観の形成促進方策の検討
MLIT-2:U	https://www.mlit.go.jp/common/001270459.pdf
MLIT-1-2:4	函館は、かつて孤立した島であった函館山が砂の堆積によって陸地と結びつき、ほかに類のない独特の地形を生み出し、函館山の裾野から北へ扇状に広がる市街地が形成された。
MLIT-2:53	・函館市の重要な観光資源の１つである函館山の夜間景観を保全、活用するため、必要に応じて周辺の市町村との調整や観光・イベントとの連携等に取り組みことが想定できる。
MLIT-2:56	市では，このような函館らしい歴史と文化を形づくっている景観を有している地域を「都市景観形成地域」に指定して，歴史的な景観の保全に努めています。
MLIT-3:T	「旧奈良監獄」を上質な「文化財ホテル」として整備～旧奈良監獄の保存及び活用に係る公共施設等運営事業を国土交通大臣が認定～
MLIT-3:U	https://www.mlit.go.jp/report/press/content/001472428.pdf
MLIT-4-1:T	道の記憶　奥州街道-その１　青森県への入り口は大難所・簑ケ坂
MLIT-4-1:U	http://www.thr.mlit.go.jp/aomori/study/oosyuu/index.html
MLIT-4-2:T	道の記憶　奥州街道-その２　南部氏の古い城跡が残る三戸町
MLIT-4-2:U	http://www.thr.mlit.go.jp/aomori/study/oosyuu/9/p11.html
MLIT-4-2:4	南部氏の居城は後に福岡城、盛岡城と移り、寛永 10 年（1633）三戸城は廃城とされ、奥州街道の時代には「三戸古城」と呼ばれていた。
MLIT-4-2:5	この城跡は現在城山公園として整備され、県南随一の桜の名所となっている。
MLIT-4-2:20	二日町から六日町へ至る街道に、「黄金橋」が架かっている。この橋にはいわれが残り、南部 12 代政行（1388 年没）の詠んだ歌が後村上天皇に賞賛され、褒美に京都加茂川の橋の擬宝珠を模することを許されたという。
JTA-1:T	第六章　災害からの観光復興
JTA-1:U	https://www.mlit.go.jp/common/001237082.pdf
TLDB-1:T	三陸の歴史
TLDB-1:U	https://www.pa.thr.mlit.go.jp/kamaishi/040/040/20200101012000.html
TLDB-1:9	古生物群については、久慈地方の琥珀や、岩泉町小本の茂師海岸で発見されたモシ竜の化石、宮古層群等で発見されるアンモナイトの化石、大船渡市日頃市町でみつかる三葉虫および四放サンゴを始め、多くの化石からも太古の三陸海岸の様子をうかがい知ることができます。
TLDB-1:25	平成 4 年、三陸・海の博覧会の折に船大工によって復元建造された 350 石（約 60 トン）積みの気仙丸。
TLDB-1:60	中世から江戸時代にかけての釜石は、東北の小さな一漁村にすぎませんでした。
TLDB-1:63	明治 7 年には官営釜石製鉄所が発足し、さらに明治 13 年には釜石と大橋の間に日本で 3 番目の鉄道が敷設され、海辺の寒村がまさに産業革命でも起きたような賑わいを見せ始めます。
TLDB-1:71	昭和 45 年の釜石。この年、八幡製鉄と富士製鉄が合併、新日本製鉄となる。
TLDB-1:74	この地図には気仙郡唐丹村（現釜石市唐丹町）に始まり、平田村、嬉石浜、釜石村、両石村、三貫島、片岸村、大槌村、吉里吉里村、船越村、織笠村、飯岡村、山田村、大沢浜、ブナ峠、豊間根村、重茂村、津軽石村、金ヶ浜、高浜、小袖浦、磯鶏村、宮古、鍬ヶ崎、崎山村、田老村（現宮古市田老）など沿岸の村々が明記されている。

表 8　アノテーションに用いた記事の情報（6～12 記事目）