UD Japanese GSD の再整備と固有表現情報付与

松田 寛 * 若狭 絢 山下 華代 大村 舞 浅原 正幸
Megagon Labs, Tokyo 国立国語研究所 フリー 国立国語研究所 国立国語研究所
株式会社リクルート

1. はじめに

Universal Dependencies (UD) は、多言語間で共通のアノテーション方式を用いて係り受けのツリーバンクを開発する国際プロジェクトである。浅原ほか (2019) は日本語の UD リソースの現状をまとめている。この中で、元テキストも含めて再配布可能なものは UD Japanese PUD と UD Japanese GSD の 2 つである。しかしながら、これらのリソースも、ライセンスや文の欠損などのさまざまな問題を抱えている。これらの問題を解決するために、我々は UD Japanese GSD の再整備を進めている。ライセンスや失われた情報の復元を進めるとともに、ほかの日本語 UD リソースに合わせて、Omura and Asahara (2018) の手法に基づいたデータの整備を進めた。また、新たに固有表現情報を付与した。

これらの作業により、spaCy 標準日本語モデルへの依存構造解析・固有表現抽出モデルの搭載が可能になる。spaCy は多言語の字句解析・固有表現抽出・品詞タグ付け・ラベル付き依存構造解析機能を提供する汎用自然言語処理フレームワークであるが、言語モデル整備時に学習元データを同梱する必要があった。今回、商用利用可能なライセンスに変更し、固有表現情報を付与した UD Japanese GSD を再整備することで spaCy の言語モデル整備に必要な標準的な要件を満たすことになる。

2. UD Japanese GSD 再整備作業

ライセンスの修正： UD Japanese GSD は元テキストとして wikipedia のテキストを含んでいる。wikipedia は CC BY-SA 3.0 もしくは GFDL に基づく文章素材の二次利用を許している。しかしながら、開発の過程でライセンスが CC BY-NC-SA に変更されていた。過去の開発者に確認したうえで CC BY-SA に変更した。

データの復元作業： UD Japanese GSD は過去の整備において、さまざまな情報が失われていた。係り受け解析器を構成するために不都合な文や句が失われており、文意がとれないテキストが確認されている。検索などで元テキストをあたりながら失われた文や句の復元をつとめた。また、英単語などにおける前後の空白の情報も失われていた (例 ”New York” → ”NewYork”)。全データを確認し、単語間に空白がある場合には、CoNLL-U Format の 10 列目の MISC の列に SpaceAfter=True を付与する作業を行った。

文節係り受け情報付与： UD Japanese GSD の元テキストに対して、浅原・松本 (2018) 互換の文節係り受け情報を付与した。整備にあたっては、国語研短単位形態論情報 (小椋ほか 2011a) を人手で付与し、文節に相当する国語研長単位形態論情報 (小椋ほか 2011b) を人手で付与したうえで、文節係り受け情報を付与した。これらは形態素解析器 MeCab + UniDic と文節係り受け解析器 CaboCha と互換性のある形式で整備されているため、再配布可能な文節係り受けアノテーションデータとしても利用できる。なお、並列構造についてはアノテーションを行わなかった。

文節係り受けから UD への変換： 文節係り受けから UD への変換は Omura and Asahara (2018) と同じプログラムにより行なう。UD 全体の方針が変わった場合には、変換プログラムの改変により吸収するため、松田ほか (2019) の日本語依存構造解析器 GiNZA の学習元ファイルである UD Japanese BCCWJ と同期してメンテナンスができるようになる。

* hiroshi_matsuda@megagon.ai

基礎統計：表 1 に UD Japanese GSD の基礎統計を示す。随時、脱語・脱文の復元を進めているために、最終的な単語数・文節数・文数は変更される可能性がある。

表 1　データの基礎統計

データセット	単語数	文節数	文数
UD Japanese GSD TEST	15,333	5,253	556
UD Japanese GSD DEV	12,573	4,716	510
UD Japanese GSD TRAIN	175,481	65,696	7,158
合計	203,387	75,765	8,224

3.　固有表現情報付与作業

固有表現ラベルの定義： 本研究では Weischedel et al. (2013) の OntoNotes5 (OntoNotes Release 5.0) の固有表現ラベル体系をベースに用いた。これは、同じく OntoNotes5 の固有表現ラベル体系を使用する spaCy での利用を容易にするためである。本作業で用いた固有表現ラベル体系を表 2 に示す。

表 2　OntoNotes Release 5.0 固有表現ラベル体系と追加定義したラベル

PERSON	People, including fictional.
NORP	Nationalities or religious or political groups.
LANGUAGE	Any named language.
GPE	Countries, cities, states.
LOC	Non-GPE locations, mountain ranges, bodies of water.
EVENT	Named hurricanes, battles, wars, sports events, etc.
FAC	Buildings, airports, highways, bridges, etc.
ORG	Companies, agencies, institutions, etc.
PRODUCT	Objects, vehicles, foods, etc. (Not services.)
WORK_OF_ART	Titles of books, songs, etc.
LAW	Named documents made into laws.
DATE	Absolute or relative dates or periods.
TIME	Times smaller than a day.
PERCENT	Percentage, including ”%”.
MONEY	Monetary values, including unit.
QUANTITY	Measurements, as of weight or distance.
ORDINAL	“first”, “second”, etc.
CARDINAL	Numerals that do not fall under another type.
※以下は追加項目	
PHONE	Phone numbers.
EMAIL	Email addresses.
URL	URLs.
PET_NAME	Individual animal names, including fictional.

関根の拡張固有表現階層との対応付け： OntoNotes5 は英語・中国語・アラビア語のデータセットであり、日本語を対象とした固有表現ラベルの定義は存在しないため、OntoNotes5 の固有表現ラベル体系を日本語に適用するための基準を策定した。固有表現ラベル付与作業の一貫性を確保する上で重要な語句およびスパンの認定基準として、関根ほか (2018) の関根の拡張固有表現階層バージョン 8（一部にバージョン 7 独自の項目を含む）の定義、および、同体系で固有表現ラベルが付与された橋本ほか (2008) の GSK2014-A の事例を参照し、表 3 のように OntoNotes5 の固有表現ラベルと関根の拡張固有表現階層の各エントリを対応づけた。

関根の拡張固有表現階層の全エントリのうち、OntoNotes5 に対応する固有表現ラベルが存在したものは 215 件あった。OntoNotes5 に対応する固有表現ラベルが存在しないエントリのうち、産業応用において重要と考えられる Phone_Number・Email・URL については、それぞれ対応するラベルを追加で定義した。また、ペットや競走馬など人以外の生物や擬人化された個体につけられた名称に対するラベルとして PET_NAME を追加で定義した。最終的に、OntoNotes5 や独自に追加定義した固有表現ラベルにマッピングされなかった関根の拡張固有表現階層のエントリは 56 件あった。その多くは物質や生物などの自然物名となっている。

表 3　本研究で使用した固有表現ラベル体系と関根の拡張固有表現階層の対応

本研究のラベル	対応する関根の拡張固有表現階層のエントリ（一部にバージョン 7 のものを含む）
PERSON	Person, God
NORP	International_Organization, Ethnic_Group, Ethnic_Group_Other, Nationality, Political_Organization, Political_Organization_Other, Political_Party, Religion
LANGUAGE	Language, Language_Other, National_Language
LOC	Location_Other, GPE, GPE_Other, City, Province, Country, Spa, Address, Address_Other, Postal_Address, County
GPE	Region, Region_Other, Continental_Region, Domestic_Region, Geological_Region, Geological_Region_Other, Mountain, Island, River, Lake, Sea, Bay
EVENT	Event_Other, Occasion, Occasion_Other, Election, Religious_Festival, Competition, Game, Conference, Incident, Incident_Other, War, Natural_Phenomenon, Natural_Phenomenon_Other, Natural_Disaster, Earthquake
FAC	Facility, Facility_Other, Facility_Part, Dam, Archaeological_Place, Archaeological_Place_Other, Tomb, FOE, FOE_Other, GOE_Other, Military_Base, Power_Plant, Park, Shopping_Complex, Sports_Facility, Museum, Zoo, Amusement_Park, Theater, Worship_Place, Castle, Palace, Public_Institution, Accommodation, Medical_Institution, School, Research_Institute, Market, Transport_Facility, Transport_Facility_Other, Car_Stop, Station, Airport, Port, Line, Line_Other, Railroad, Road, Canal, Water_Route, Tunnel, Bridge, Tumulus
ORG	Organization, Organization_Other, Show_Organization, Family, Sports_Organization, Sports_Organization_Other, Pro_Sports_Organization, Sports_Federation, Sports_League, Sports_Team, Juridical_Person, Juridical_Person_Other, Channel, Corporation_Other, Nonprofit_Organization, Company, Company_Group, Government, Cabinet, Military
PRODUCT	Product_Other, Service, Character, ID_Number, Game_Other, Degital_Game, Software, Vehicle, Vehicle_Other, Car, Train, Aircraft, Spaceship, Ship, Food_Other, Musical_Instrument, Clothing, Money_Form, Drug, Weapon, Stock, Award, Decoration
WORK_OF_ART	Video_Work, Art, Art_Other, Painting, Broadcast_Program, Movie, Show, Music, Book, Printing, Printing_Other, Newspaper, Magazine, Picture
LAW	Offense, Doctrine_Method_Other, Movement, Plan, Rule, Rule_Other, Treaty, Law
DATE	Timex, Timex_Other, Timeex, Timeex_Other, Date, Day_Of_Week, Era, Periodx, Periodx_Other, Period_Day, Period_Week, Period_Month, Period_Year, Time_Top_Other
TIME	Time, Period_Time
PERCENT	Percent
MONEY	Currency, Money
QUANTITY	Unit_Other, Latitude_Longitude, Latitude_Longtitude, Measurement, Measurement_Other, Physical_Extent, Seismic_Magnitude, Space, Volume, Weight, Speed, Intensity, Temperature, Calorie, Seismic_Intensity, Countx, Countx_Other, N_Person, N_Organization, N_Location, N_Location_Other, N_Country, N_Facility, N_Product, N_Event, N_Natural_Object, N_Natural_Object_Other, N_Animal, N_Flora, Point, Multiplication, Frequency, Age
ORDINAL	Rank, School_Age, Ordinal_Number
CARDINAL	Stock_Index
※以下は追加項目	
PHONE	Phone_Number
EMAIL	Email
URL	URL
PET_NAME	Individual_Animal, Individual_Animal_Other, Racehorse

4. 成果の応用

本研究で再整備された UD Japanese GSD でパラメータを学習した解析モデルは商用利用も可能であり、UD Japanese GSD を学習コーパスとして使用している Stanford NLP など多くの汎用自然言語処理フレームワークの日本語モデルの利用が促進されると考えている。これらのフレームワークの多くは、字句解析を含む全ての解析処理を深層学習技術を用いて学習する end-to-end のアプローチを取っている。そのため、現状の学習コーパスの規模では十分な語彙を学習できず、日本語の解析で用いられてきた数十万語彙を備える形態素解析器を前段に用いるアプローチと比較してトークン化の精度が低く、速度もかなり遅い傾向にある。

一方、産業応用を念頭に開発された spaCy の標準日本語モデルは、速度と精度の両面から MeCab や SudachiPy などの形態素解析器を用いて実装が行われてきたが、依存構造解析および固有表現抽出モデルは適切な学習コーパスが存在しなかったため（GiNZA のような派生ライブラリを除いて）公式には提供されてこなかった。本研究で UD Japanese GSD が再整備され固有表現情報も追加されたことにより、spaCy の特長である依存構造解析と固有表現抽出のマルチタスク学習が可能となった。今後、spaCy の GitHub リポジトリの Issue を通じて、UD Japanese GSD で学習した標準日本語モデルの公開を推進する予定である。

なお、本研究の固有表現情報付与作業の候補抽出に用いた固有表現抽出モデル（UD Japanese BCCWJ から新聞記事を除外したものと GSK2014-A のアライメントを取り spaCy でマルチタスク学習）は GiNZA version 3 に組み込んだ形で Megagon Labs の GitHub リポジトリおよび PyPI から公開している。

謝　辞

本研究は Megagon Labs と国立国語研究所の共同研究協定・国立国語研究所コーパス開発センター共同研究プロジェクトおよび科研費 JP17H00917, JP18H05521 によるものです。

文　献

浅原正幸・金山博・宮尾祐介・田中貴秋・大村舞・村脇有吾・松本裕治 (2019).「Universal Dependencies 日本語コーパス」　自然言語処理, 26:1, pp. 3–36.

Mai Omura, and Masayuki Asahara (2018). “UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese.” *Proceedings of the Second Workshop on Universal Dependencies*, pp. 117–125. Brussels, Belgium: Association for Computational Linguistics.

浅原正幸・松本裕治 (2018).「『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造アノテーション」　自然言語処理, 25:4, pp. 331–356.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011a).「『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（下）」　Technical report, 国立国語研究所. JC-D-10-05-02

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011b).「『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上）」　Technical report, 国立国語研究所. JC-D-10-05-01

松田寛・大村舞・浅原正幸 (2019).「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.

Ralph Weischedel et al. (2013). *OntoNotes Release 5.0 LDC2013T19*. Philadelphia, PA: Linguistic Data Consortium.

関根聡・安藤まや・小林暁雄・松田耕史・鈴木正敏・Duc Nguyen・乾健太郎 (2018).「「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.

橋本泰一・乾孝司・村上浩司 (2008).「拡張固有表現タグ付きコーパスの構築」　情報処理学会研究報告, 自然言語処理研究会報告 (NL-188-17), pp. 113–120.