日本語固有表現アノテーションガイドライン

Version 1.0

前川恵美

2019/09/17

1. 概要

本仕様書で用いるタグセットは、「OntoNotes Release 5.0」[1]にて用いられている 18 種類と新たに設定した 5 種類のカテゴリで構成される。各タグの詳細な定義は、原則として「OntoNotes Named Entity Guidelines Version 14.0」[2]に準拠した。

2. タグセット一覧

各タグは、以下の各タグの先頭に NE-を付与したものとする。例：NE-PER

タグ	OntoNote におけるタグ	内容	説明	例　該当部分を( )で囲む
PER	PERSON	人名	敬称等は含まない。但し「～世」「女王」「皇后」「聖」等は含む。 動物の愛称は含まない。人間に対するキャラクター名、あだなは含む。 神なども含む。	(ヨハネス三世) (佐藤)教授 / (エリザベス女王) (田中)首相 / (まっちゃん) (チャーリー・ブラウン) (スサノオノミコト)
PER-N	新設	地位・職業名	敬称等（様、先生等）は含まない。普通名詞は含まない。 役職名の前の修飾語句・接頭辞「元」「前」「旧」などは含まない。	ブッシュ(米大統領) (マイクロソフトＣＥＯ) 前(運輸大臣) 佐藤(裁判長)
CHA	新設	人間以外の生き物名	人間以外の生き物の名前。動物。神・女神は PER。	(スヌーピー) / (ポチ)
NRP	NORP	国籍/宗教/政治的属性	国籍、宗教・政治グループの構成員。 英語形容詞のカタカナ化。 直後に人間を示す言葉が続く国・地域名。 明らかに「～的な／風の」という意味で使われている国・地域名。 国・地域名に「～式」「～的」「～風」「～性」が続くもの（「式」まで含む）。	(日本人) / (仏教徒) / (共和党員) (イタリアン)レストラン (米)兵 / (仏)観光客 (日本)庭園 / (フランス)料理 (地中海性)高気圧 (広島風)お好み焼き (ニューヨーカー) / (京女)
FAC	FACILITY	建造物	人工的に作られた建造物。ビル、空港、道路、橋、モニュメント等。空港内の施設。通りの名前。油田、駅、ゴルフコース、病院、動物園等。 文脈により ORG になる。	(ブルックリン・ブリッジ) (甲子園球場) / (六本木ヒルズ) (阪神高速道路) / (河原町通) (新大阪駅) / (自由の女神) (NE-FAC 中之島図書館)の外壁に… (NE-ORG 中之島図書館)の発表では

ADD	新設	連絡先	住所、電話番号、URL、メールアドレス。州/県＋市町村区名等(例：東京都渋谷区)は ADD ではなく、それぞれを FAC や GPE とする。	(神戸市中央区中山手通り 10-2-1) 「神戸市中央区中山手通り」だけなら(NE-GPE 神戸市)(NE-GPE 中央区)(NE-GPE 中山手通り)
ORG	ORGANIZA-TION	組織	会社、政府、行政組織、政党名、教育組織、チーム。ホワイトハウスも ORG。文脈により FAC にもなりうる。特にホテル、美術館、病院、図書館、寺院、商店、証券取引所などは FAC になることが多い。 新聞・雑誌・ウエブサイトの名前は文脈に関係なく常に ORG。 都市名などだけでスポーツチームを表しているものも ORG。	(IBM) / (日本共産党) / (外務省) (タイガース) / (ホワイトハウス) (キャピトル・ヒル) (ヒルトンホテル) (ニューヨーク証券取引所) (毎日新聞) / (週刊新潮) / (YouTube) (ウィキペディア)の記事 (日本政府) / (オバマ政権) / (米軍) c.f. (NE-FAC ニューヨーク証券取引所)の外では雪が降っていた。
GPE	GPE	国/市町村/州等	行政単位であること。島の名前も行政単位なら GPE。GPE か LOC かわからない場合は GPE とする。	(米国)企業 / (世田谷区) / (インド) (マンハッタン) / (大阪府) (大阪)北部
LOC	LOCATION	GPE 以外の地名	天体、星、大陸、山、海、海岸、河川、湖、国境など。ある地域（中東、ミナミ、Soho など）。	(淀川) / (ヨーロッパ) / (アフリカ) (須磨海岸) / (太平洋) / (土星) (ミナミ) / (イーストビレッジ) (三宮) / (紀伊半島)
PRO	PRODUCT	製品	商品でない乗り物名（車、ロケット、航空機、船）も含む。 メーカー名のみで製品を指しているときは ORG ではなく PRO。 サービス・金融製品・クレジットカードも含む(OntoNotes とは異なる)。	((NE-ORG トヨタ) (NE-PRO プリウス)) ２台の(NE-PRO ポルシェ) (スペースシャトル) (ドリームジャンボ) / (正露丸) (Windows10) / (iPhone) (ヒートテック)
EVT	EVENT	現象・出来事	事故・災害、天変地異、革命、戦争、スポーツイベントなどを含むイベント。	(ハリケーン・カトリーヌ) (阪神淡路大震災) / (ベトナム戦争) (文革) / (応仁の乱) / (９・１１) (2016 年ワールドカップ) / (F1)
ART	WORK OF ART	タイトル	知的生産物のタイトル。賞の名称。株価指数。年金・保険等のシステム名。新聞の見出しは参照されている場合は含むが、新聞として使用されている場合は含まない。作品のタイトルそのものではないシリーズ名も含む。	(吾輩は猫である) / (イエスタデイ) (オスカー) / (ノーベル賞) (日経平均株価) / (国民健康保険) (ユニオンジャック) (NE-ART サザエさん)を TV で見る。 (NE-PER サザエ)さんは買い物に。

PJT	新設	プロジェクト	政策、プロジェクト、計画、運動、主義、宗教、流派、システムなど。文脈により ORG にもなる。	(ニューディール) (ペレストロイカ) (キリスト教)
LAW	LAW	法	法律、条令、憲法、条約など。	(権利章典) / (憲法第九条) (ワルシャワ条約) (労働基準法第二条第一号)
LAN	LANGUAGE	言語	プログラミング言語も含める。	(日本語) / (C言語)
DT	DATE	日以上の時	日、曜日、月、ある一定の期間、季節、「今日」「来年」。 年齢（数だけの場合も。形容詞や副詞も）。 「最近」「現代」「かつて」「昔」「将来」「未来」等は含まない。 基本的に 24 時間以上は DT。 「～前」「～以降」なども含める。 範囲を示す表現は全体を一つの DT とする。助詞の「の」も含める。	(2001 年 12 月１日) / (月曜日) (数年) / (今日) / (明日) / (子供の日) (毎年) / (ここ数ヶ月) / (過去３日) (先週) / 山田太郎(５１)が… (上半期) / (春) / (夏期)休暇 (１年前) / (５月から９月まで) (５０代) / (江戸時代) (１月３、４、５日) (２１世紀) / (４８時間) (昨年の１２月)
TM	TIME	日未満の時	時刻、時間。24 時間は DT。 範囲を示す表現は全体を一つの TM とする。助詞の「の」も含める。	(午後４時) / (今朝) / (３時間) / (１秒) (午前中) / (今日の２時) (９時から５時まで) c.f. (NE-DT ２４時間)
PC	PERCENT	パーセント	「%」まで含む。「半分」は含まない。	(１０%) / (８割) / (ほぼ１００パー) c.f. (NE-CD 半分)
MO	MONEY	金額	単位を伴っていること。 株式で使用する「1 株あたり」などは含まない。	(百円) / (５００米ドル) (US$３００) / (１ドル２５セント) １株あたり(３円９０銭)
QT	QUANTITY	数量	単位を伴っていること。長さ、距離、面積、熱量、速度、温度、重量、バイト数など	(約５キロ) / (４００m)リレー (十坪以上) / (時速８０キロ) (５メガバイト)
OD	ORDINAL	序数	序数詞。物事の順序を示す名詞。副詞的名詞も含む。 「最初」は含むが「最後」は含まない。 助詞「の」や助動詞「だ」等は含めない。	(２番目) / (三段目) / (４つめ) (２回目) / (３位) / (第二)の (一次)選考 / (３度目) (最初)だ
CD2	新設	数＋単位	数量表現 QT 以外の「数＋単位」。 「単位」は、何かを数える語としてのみ使用される語であること。それ以外	(１００人) / (五台) / (３つ) (数百万頭) / (６００冊) / (６校) (２ヶ国) / (二度) / (７回) c.f.(７)回表

			は含まない。 回数表現「一度」「一回」は含める。 野球のイニング「一回」は含めない。	 (NE-CD ２)点 / (NE-CD ５０)店舗 (NE-CD ３)往復 / (NE-CD ４)回転
CD	CARDINAL	上記以外の数	単位を伴わない､上記のどのカテゴリにも属さない数。 リスト等の項目を示す数字。 「約」「ほぼ」「以上」等も含む。	(半分) / (数百) / (３分の２) (５)対(０) / (１～３)大学 (２)点 / (５０)店舗 / (３)往復 (４)回転 / (７)回表 c.f. (７回)読む
UC	新設	未分類	上記どのカテゴリにも属さないもの、或いはどこに分類してよいかよくわからないもの。	(アルツハイマー)病

3. NE とする範囲

■句読点や助詞を含むもの

日付、時間、数量表現以外で、句読点や助詞を含む形が正式名称である場合(例 : 道の駅 / モーニング娘。)を除いて、句読点や助詞で区切られた部分をそれぞれ個別の NE とする :

神戸市の中央区で　→　(NE-GPE 神戸市)の(NE-GPE 中央区)で

日付・時間・数量は句読点や助詞で区切られていても、その時や期日を示す最大範囲を一つの NE とする。

８月末日の午後２時に　→　(NE-TM ８月末日の午後２時)に

ただし、構造上異なる階層に属する要素は別々にタグ付けする。

月曜、午後３時、公演が始まった。
→ (NE-DT 月曜)、(NE-TM 午後３時)、公演が始まった。
＊(NE-DT 月曜、午後３時) とはならない。

また、範囲を示す「…から…まで」や「…間」は NE の範囲内に含める。

４月８日から２０日まで → (NE-DT ４月８日から２０日まで)　c.f. 期間は(NE-DT ４月８日)まで
４月の８日から２０日まで → (NE-DT ４月の８日から２０日まで)
２０日間 → (NE-DT ２０日間)

ただし、「間」の前に助詞の「の」が入っている場合には「の間」は範囲に含めない。

２０日の間 → (NE-DT ２０日)の間

■かっこつき : 同格

かっこつきの挿入句が直前の名詞と同格である場合は、かっこの外と中のものにそれぞれ別のタグを付与する。

アジア太平洋経済協力会議（ＡＰＥＣ）→(NE-ORG アジア太平洋経済協力会議) （ (NE-ORGＡＰＥＣ) ）

3550 ポンド（16100 キロ）→ (NE-QT 3550 ポンド) ((NE-QT 16100 キロ) ）

ただし、以下のように途中に挿入されている場合は全体を一つの NE とする（英語仕様と異なる）:

火曜日（１月４日）の１２時 → （NE-TM 火曜日（１月４日）の１２時）

挿入でない場合でも、略称や言い換えなどが（　）つきで表示されている場合に係り受けの関係から切れない場合にはそれも含める。

１００キロ（ｋｍ）→ (NE-QT １００キロ（ｋｍ）)
係り受けで「キロ」と「(ｋｍ)」が先に要素化されており、１００キロと（ｋｍ）分割できない。

例外：お金を異なる通貨で言い換えている場合は、挿入であっても、切る

１０億ドル（７億ユーロ）以上 → (NE-MO １０億ドル)（(NE-MO ７億ユーロ)）以上

■かっこつき：補足

時間・数量表現について、補足情報などをかっこ付きで表現している場合、かっこで囲まれた部分はその時間・数量表現に含める。（英語仕様と異なる）

午後２時（東部標準時） → (NE-TM 午後２時（東部標準時）)

4. 追加カテゴリ

■PER-N：地位・職業名

固有名詞として出現している役職や職業名。
大文字表記のある英語とは異なり、普通名詞と固有名詞の判別が困難な場合もあるので、以下の 3 つの形を基本的には PER-N とする：

- **役職名＋名前**：ロシアの大統領、プーチン → ロシアの(NE-PER-N 大統領/NN)、プーチン
- **名前＋役職名**：プーチン大統領 → プーチン(NE-PER-N 大統領/NNP)
- **役職名＋「の」＋名前**：大統領のプーチン → (NE-PER-N 大統領/NN)のプーチン

先生、教授、様等の敬称は PER-N とはしない。迷ったら PER-N とする

「大統領」や「CEO」などが文脈上誰のことを指しているか明白であっても、それのみで登場した場合には PER-N とはしない。

プーチン大統領が退院。**大統領**は… → (NE-PER プーチン)(NE-PER-N 大統領)が退院。**大統領**は…

ただし、その前に組織名などが付与されており、指している対象を特定した表現の場合は PER-N とする。

(NE-PER マイクロソフト CEO)　(NE-PER 米大統領)

■CHA：人間以外の生物の名前

人間以外の生物の名称。動物など。人間の形をした妖精や神などは PER として扱う。

(NE-CHA のらくろ)　(NE-CHA キティ)

■ADD：連絡先情報

住所、電話番号、電子メールアドレス、URL など。
住所は OntoNote によるガイドラインでは個別の NE に分けるとされているが、本仕様は ADD タグを追

加、ひとまとめのものとして NE-ADD とする。尚、住所を構成するそれぞれの要素が単独（或いは住所としては不完全な形）で出現する場合には ADD とはせず、それぞれ適切なタグをふる。

神戸市中央区三宮町１０－２ →(NE-ADD 神戸市中央区三宮町１０－２)
御堂筋にある →(NE-FAC 御堂筋) にある
中央区三宮町 → (NE-GPE 中央区) (NE-GPE 三宮町)

不完全な形の住所であっても、番地＋町名(ストリート名など)の場合は ADD とする。

三宮町１０－２ →(NE-ADD 三宮町１０－２)

URL はそのままウェブサイト名となる場合が多い。サイト名として扱われている場合には ORG とする。

■PJT：プロジェクト

プロジェクト、プログラム等以外に新興宗教や思想なども。英語ではいわゆる四大宗教(キリスト教他)は PJT としなかったが、日本語では四大宗教も含める。

■UC：未分類

どのカテゴリにも属さない固有表現、或いは、アノテータがどこに分類してよいかわからない固有表現。

5. 判断に迷う場合の処理方法

■主辞を共有している場合

例：兵庫、岡山県

OntoNotes では、(NE-GPE　兵庫、岡山県)とはせずに２つの別々の固有表現とみなし、主辞に最も近い要素のみを NE として採用する（つまり兵庫、(NE-GPE 岡山県)）としているが、PTB の構造では等位接続詞や読点でつながれた部分が先に一つの階層を形成するための兵庫、岡山県部分を分割することはできない。よって、このような場合は全体で１つの NE とする：

(NE-GPE　兵庫、岡山県)

日付 DT と時間 TM の場合も同様。

７月８・９日 →　(NE-DT　７月８・９日)

ただし、間に接続詞や読点が入っており、個別に NE としてとっても（表現として）そのものだとわかる場合には個別にタグをふる。

ラガーディア、ジョン・Ｆ・ケネディおよびニューアーク国際空港
→ (NE-FAC ラガーディア)、(NE-FAC ジョン・Ｆ・ケネディ)および(NE-FAC ニューアーク国際空港)

■地名：LOC か GPE か

地名（特に海外）が市町村、州、県などの行政単位であるかどうか不明の場合は GPE とする。

■NRP と GPE／LOC について

GPE や LOC となる名称は、以下の場合 NRP とする：

・「～式」「～的」「～風」「～性」

(NE-NRP アメリカ的)暮らし　(NE-NRP 日本式)家屋　(NE-NRP 京風)ラーメン
(NE-NRP 地中海性)高気圧

・明らかに「〜的な／風の」という意味で使われているとき（国家や地理的な存在ではない）

(NE-NRP 日本)庭園　　(NE-NRP イタリア)料理

・英語形容詞のカタカナ化

(NE-NRP イタリアン)レストラン　(NE-NRP フレンチ)ギャル　　(NE-NRP アメリカン)スタイル

・国籍など、あるエリアの居住者

(NE-NRP フランス人)　(NE-NRP 京女)　(NE-NRP ニューヨーカー)　　c.f. (NE-GPE フランス)の人

・直後に人間を示す言葉が続く国・地域名

(NE-NRP ポーランド)観光客（＊ポーランドへの観光客ではなく、ポーランド人の観光客という意味で）
(NE-NRP 米)兵

6. 参考文献

[1] Ralph Weischedel 他（2012）「*OntoNotes Release 5.0*」[online]
https://catalog.ldc.upenn.edu/docs/LDC2013T19/OntoNotes-Release-5.0.pdf（参照 2017 年 6 月 1 日）

[2] Raytheon BBN Technologies「*OntoNotes Named Entity Guidelines Version 14.0.*」 [online]
https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxjbHRsYW5ub3RhdGlvbnN8Z3g6MzVlOGFkMjQ4ZWMxM2Y5MA（参照 2017 年 7 月 1 日）