P1-34 UD Japanese GSDの再整備と固有表現情報付与

（言語処理学会 第26回年次大会 発表資料）

松田 寛　Megagon Labs
若狭 絢　国立国語研究所
山下 華代　フリー
大村 舞　国立国語研究所
浅原 正幸　国立国語研究所
㈱リクルート

● 取り組みの背景

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ツリーバンクを開発する国際プロジェクトである。日本語のUDコーパスのうち元テキストを含めて再配布可能なものは、全言語共通の1000文(翻訳)で構成されるPUDと、Wikipedia等のテキスト約8000文で構成されるGSDの2種類がある。 ただしこれらのコーパスから依存構造解析モデルを学習するには障壁があった。

● 誰もが自由に使えるコーパスを目指して

日本語の係り受け情報つきコーパスで、明に**商用利用可能なライセンス**のもとで公開されたものはなく、世界的に普及が進む**各種NLPフレームワークでの日本語モデル利用の障壁**となっている。そこで既存のUD Japanese GSDコーパスを再整備し、**CC BY-SA ライセンスでの公開**を目指した。同時に品質改善に取り組みながら、**一般的な形態素解析器を適用可能**な形で係り受け・固有表現情報の人手付与を行った。

テキストの再整備と文節係り受け情報の付与

● ライセンスの修正

GSDはWikipediaのテキストを含みCC BY-SA 3.0(もしくはGFDL) で配布する必要があるが、開発過程でCC BY-NC-SAに変更されていたため、関係者に確認の上で**商用利用可能なCC BY-SAに変更**した。

● データの復元作業

従来のGSDでは原文の一部が削除されて文意がとれないテキストがあったため、Web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**につとめた。また英単語前後の空白情報の復元も行った。

● 文節係り受け情報の人手による付与

- **国語研短単位形態論**情報の付与
- 文節相当の国語研長単位形態論情報の付与
- 文節係り受け情報の付与
- MeCab+UniDic+CaboChaと互換性のある形式

● 文節係り受けからUDへの自動変換

全ての日本語UDコーパスを同時にメンテナンス可能

- UniDic品詞からUPOSへの変換 - 約40種類
- 文節係りから単語間係りへの変換(主辞間＋文節内)
- 依存関係ラベルの付与 - 約60種類

短単位品詞	短単位基本形	長単位用法	UPOS
^形容詞-非自立可能		形容詞-一般	ADJ
^形容詞-非自立可能		助動詞	AUX
^名詞-普通名詞-サ変可能		名詞-普通名詞-一般	NOUN
^名詞-普通名詞-サ変可能		動詞-一般	VERB
^連体詞	^[こそあど此其彼] の		DET
^連体詞	^[こそあど此其彼]		PRON
^動詞-非自立可能			AUX
^動詞			VERB
^名詞-固有名詞			PROPN
^名詞-普通名詞-副詞可能		副詞	ADV
^名詞-普通名詞-副詞可能			NOUN
接頭辞			NOUN

ラベル付与ルール	UDラベル
その係り元単語は係り先がなく (文末の文節である) でさらに文節の主辞である	root
その係り元単語は UPOSNUMMOD を持っている。	nummod
その係り元単語は UPOSADV を持っている	advmod
係り先単語は VERB を持っており、格助詞「が」が文節内にある	nsubj
係り先単語は VERB を持っており、格助詞「を」が文節内にある	obj
その係り元単語は UPOSVERB 持っており、その係り先単語はUPOSVERB を持っており、文節をまたがっている	aux
その係り元単語は UPOSVERB 持っており、その係り先単語は UPOSVERB を持っており、文節内の関係である	compound

固有表現情報の付与

- 固有表現ラベルの定義

本研究ではspaCyが用いる**OntoNotes5の固有表現ラベル体系にいくつかの拡張を加えて使用**した。OntoNotes5は英語・中国語・アラビア語のデータセットであるため、OntoNotes5の固有表現ラベル体系を日本語に適用するための独自の基準を策定した。

- 関根の拡張固有表現階層との対応付け

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、関根の拡張固有表現階層の定義、および、同体系で固有表現ラベルが付与された**GSK2014-Aの事例を参照**した。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に対応が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ものを**PHONE・EMAIL・URL・PET_NAMEとして追加定義**した。

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.

Universal Dependencies体系について

● Universal Dependencies品詞体系(17種)

UniDic短単位品詞で「～可能」が末尾につく品詞は用法の曖昧性解消が必要(本稿では長単位を参照)。
例: 名詞-普通名詞-サ変可能 → NOUN or VERB

Open class words
- ADJ: adjective
- ADV: adverb
- INTJ: interjection
- NOUN: noun, PROPN: proper noun
- VERB: verb

Closed class words
- ADP: adposition
- AUX: auxiliary
- CCONJ: coordinating conjunction
- DET: determiner
- NUM: numeral
- PART: particle
- PRON: pronoun
- SCONJ: subordinating conjunction

Other
- PUNCT: punctuation, SYM: symbol, X: other

● Dependency Labels(37種)

nsubj, obj, iobj等の格関係は基本的に表層格で決定するが一部は述語項構造まで参照。

acl: clausal modifier of noun (adjectival clause)
advcl: adverbial clause modifier
advmod: adverbial modifier, amod: adjectival modifier
appos: appositional modifier, aux: auxiliary
case: case marking, cc: coordinating conjunction
ccomp: clausal complement, clf: classifier
compound: compound, conj: conjunct, cop: copula
csubj: clausal subject, dep: unspecified dependency
det: determiner, discourse: discourse element
dislocated: dislocated elements, expl: expletive
fixed: fixed multiword expression
flat: flat multiword expression, goeswith: goes with
iobj: indirect object, list: list, mark: marker
nmod: nominal modifier, nsubj: nominal subject
nummod: numeric modifier, obj: object
obl: oblique nominal, orphan: orphan, parataxis: parataxis
punct: punctuation, reparandum: overridden disfluency
root: root, vocative: vocative
xcomp: open clausal complement

作業進捗と今後の展開

- 3/13時点での一次作業の進捗状況
  - 短単位・長単位形態論/係り受け情報付与 – 100%
  - 固有表現情報付与 – 30%

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

- 成果の公開時期
  - **2020.04 – GSD/PUD *β* 版**をGitHubで公開予定
  - 2020.05 – UD次期バージョン公式リリース予定
- NLPフレームワークへのモデル提供
  - GSDから学習したモデルは再配布・商用利用可能
  - spaCy日本語モデルを公式化するプルリクを提出
- サンプル(依存木をdisplaCyで描画)

  最終ページを参照

CoNLL-Uフィールド構成

#1　ID – トークン番号
#2-3 FORM, LEMMA – 表記, 正規化形
#4-5 UPOS, XPOS – UD品詞, UniDic短単位品詞
#6　FEATS – 数詞等の詳細情報
#7　HEAD – ヘッド(依存先)のトークン番号(root=0)
#8　DEPREL – ヘッドへの依存関係種別(ラベル)
#9　DEPS – 未使用

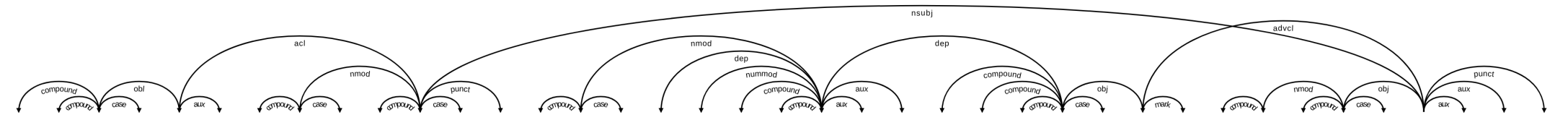
- #10=MISCフィールドの日本語拡張

BunsetuBILabel
　文節開始トークン: B, 途中のトークン: I
BunsetuPositionType
　自立語: ROOT, SEM_HEAD, CONT
　機能語: SYN_HEAD, FUNC
LUWBILabel, LUWPOS
　長単位スパンラベル(B, I)とUniDic長単位品詞
SpaceAfter – (省略時=Yes)
　Yes:直後に空白あり, No:空白なし
NE – (省略時=O)
　固有表現スパンラベル(B, I, O)-固有表現ラベル

sent_id = ccd_ud-train-20200312-504
text = 首相補佐官になった国民新党の亀井代表は、自民党から「一本釣り」した浜田総務政務官を従えて宮城県石巻市を視察した。

1	首相	首相	NOUN	名詞-普通名詞-一般	_	3	compound	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|	SpaceAfter=No
2	補佐	補佐	NOUN	名詞-普通名詞-サ変可能	_	3	compound	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=I|LUWPOS=名詞-普通名詞-一般|	SpaceAfter=No
3	官	官	NOUN	接尾辞-名詞的-一般	_	5	obl	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=名詞-普通名詞-一般|	SpaceAfter=No
4	に	に	ADP	助詞-格助詞	_	3	case	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助詞-格助詞|	SpaceAfter=No
5	なっ	なる	VERB	動詞-非自立可能	_	11	acl	_	BunsetuBILabel=B|BunsetuPositionType=SEM_HEAD|	LUWBILabel=B|LUWPOS=動詞-一般|	SpaceAfter=No
6	た	た	AUX	助動詞	_	5	aux	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助動詞|	SpaceAfter=No
7	国民	国民	NOUN	名詞-普通名詞-一般	_	8	compound	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=名詞-普通名詞-一般|	SpaceAfter=No|NE=B-NORP
8	新党	新党	NOUN	名詞-普通名詞-一般	_	11	nmod	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=名詞-普通名詞-一般|	SpaceAfter=No|NE=I-NORP
9	の	の	ADP	助詞-格助詞	_	8	case	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助詞-格助詞|	SpaceAfter=No
10	亀井	カメイ	PROPN	名詞-固有名詞-人名-姓	_	11	compound	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=名詞-固有名詞-人名-姓|	SpaceAfter=No|NE=B-PERSON
11	代表	代表	NOUN	名詞-普通名詞-サ変可能	_	36	nsubj	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=名詞-固有名詞-人名-姓|	SpaceAfter=No
12	は	は	ADP	助詞-係助詞	_	11	case	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助詞-係助詞|	SpaceAfter=No
13	、	、	PUNCT	補助記号-読点	_	11	punct	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=補助記号-読点|	SpaceAfter=No
14	自民	自民	PROPN	名詞-固有名詞-一般	_	15	compound	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=名詞-固有名詞-一般|	SpaceAfter=No|NE=B-NORP
15	党	党	NOUN	接尾辞-名詞的-一般	_	21	nmod	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=名詞-固有名詞-一般|	SpaceAfter=No|NE=I-NORP
16	から	から	ADP	助詞-格助詞	_	15	case	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助詞-格助詞|	SpaceAfter=No
17	「	「	SYM	補助記号-括弧開	_	21	dep	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=補助記号-括弧開|	SpaceAfter=No
18	一	一	NUM	名詞-数詞	_	21	nummod	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=動詞-一般|	SpaceAfter=No|NE=B-MOVEMENT
19	本	本	NOUN	接尾辞-名詞的-助数詞	_	21	compound	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=I|LUWPOS=動詞-一般|	SpaceAfter=No|NE=I-MOVEMENT
20	釣り	釣り	NOUN	名詞-普通名詞-一般	_	21	compound	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=I|LUWPOS=動詞-一般|	SpaceAfter=No|NE=I-MOVEMENT
21	」	」	SYM	補助記号-括弧閉	_	27	dep	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=動詞-一般|	SpaceAfter=No
22	し	する	AUX	動詞-非自立可能	_	21	aux	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=I|LUWPOS=動詞-一般|	SpaceAfter=No
23	た	た	AUX	助動詞	_	21	aux	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助動詞|	SpaceAfter=No
24	浜田	ハマダ	PROPN	名詞-固有名詞-人名-姓	_	27	compound	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=名詞-固有名詞-人名-姓|	SpaceAfter=No|NE=B-PERSON
25	総務	総務	NOUN	名詞-普通名詞-一般	_	27	compound	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=I|LUWPOS=名詞-固有名詞-人名-姓|	SpaceAfter=No
26	政務	政務	NOUN	名詞-普通名詞-一般	_	27	compound	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=I|LUWPOS=名詞-固有名詞-人名-姓|	SpaceAfter=No
27	官	官	NOUN	接尾辞-名詞的-一般	_	29	obj	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=名詞-固有名詞-人名-姓|	SpaceAfter=No
28	を	を	ADP	助詞-格助詞	_	27	case	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助詞-格助詞|	SpaceAfter=No
29	従え	従える	VERB	動詞-一般	_	36	advcl	_	BunsetuBILabel=B|BunsetuPositionType=SEM_HEAD|	LUWBILabel=B|LUWPOS=動詞-一般|	SpaceAfter=No
30	て	て	SCONJ	助詞-接続助詞	_	29	mark	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助詞-接続助詞|	SpaceAfter=No
31	宮城	ミヤギ	PROPN	名詞-固有名詞-地名-一般	_	32	compound	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=名詞-固有名詞-地名-一般|	SpaceAfter=No|NE=B-GPE
32	県	県	NOUN	名詞-普通名詞-一般	_	34	nmod	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=名詞-固有名詞-地名-一般|	SpaceAfter=No|NE=I-GPE
33	石巻	イシノマキ	PROPN	名詞-固有名詞-地名-一般	_	34	compound	_	BunsetuBILabel=B|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=名詞-固有名詞-地名-一般|	SpaceAfter=No|NE=I-GPE
34	市	市	NOUN	名詞-普通名詞-一般	_	36	obj	_	BunsetuBILabel=I|BunsetuPositionType=SEM_HEAD|	LUWBILabel=I|LUWPOS=名詞-固有名詞-地名-一般|	SpaceAfter=No|NE=I-GPE
35	を	を	ADP	助詞-格助詞	_	34	case	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助詞-格助詞|	SpaceAfter=No
36	視察	視察	VERB	名詞-普通名詞-サ変可能	_	0	root	_	BunsetuBILabel=B|BunsetuPositionType=ROOT|	LUWBILabel=B|LUWPOS=動詞-一般|	SpaceAfter=No
37	し	する	AUX	動詞-非自立可能	_	36	aux	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=I|LUWPOS=動詞-一般|	SpaceAfter=No
38	た	た	AUX	助動詞	_	36	aux	_	BunsetuBILabel=I|BunsetuPositionType=SYN_HEAD|	LUWBILabel=B|LUWPOS=助動詞|	SpaceAfter=No
39	。	。	PUNCT	補助記号-句点	_	36	punct	_	BunsetuBILabel=I|BunsetuPositionType=CONT|	LUWBILabel=B|LUWPOS=補助記号-句点|	SpaceAfter=No

<--- CoNLL-Uフィールド間はタブ区切り ---> <---MISCのサブフィールド間は｜区切り(サブフィールド間のタブは実際には存在しない)--->


nsubj
advcl
acl
nmod
dep
dep
nmod
punct
compound
obl
nmod
punct
nummod
compound
aux
compound
compound
obj
punct
compound
case
aux
case
compound
case
compound
compound
aux
compound
case
mark
compound
case
aux
首相 補佐 官 に なっ た 国民 新党 の 亀井 代表 は 、 自民 党 から 「 一 本 釣り 」 し た 浜田 総務 政務 官 を 従え て 宮城 県 石巻 市 を 視察 し た 。
- - - - - - B-NORP I-NORP - I-PERSON - - - B-NORP I-NORP - - B-MOVE. I-MOVE. I-MOVE. - - - B-PERSON - - - - - - B-GPE I-GPE I-GPE I-GPE - - - - -
NOUN NOUN NOUN ADP VERB AUX NOUN NOUN ADP PROPN NOUN ADP PUNCT PROPN NOUN ADP SYM NUM NOUN NOUN SYM AUX AUX PROPN NOUN NOUN NOUN ADP VERB SCONJ PROPN NOUN PROPN NOUN ADP VERB AUX AUX PUNCT