日本語UD v2: 公開の経緯と反省点

Jun 16, 2018

Hiroshi Kanayama

hkana@jp.ibm.com

日本IBM東京基礎研究所

自己紹介：金山 博

- 日本アイ・ビー・エム株式会社 東京基礎研究所
  Knowledge Infrastructure Group
- 専門分野
  - 自然言語処理の基礎（構文解析・意味解析）
  - テキストマイニングの応用・多言語化
- 電子情報通信学会 言語理解とコミュニケーション専門研究委員長
- 言語処理学会理事

- UDに関する活動
  - 2015年言語処理学会年次大会での発表
  - 2016年 UDコアチームとの議論
  - 2017年言語処理学会年次大会でのチュートリアル（NTT田中さんと）
  - 2017年～2018年 Japanese GSD/PUDデータの作成

発表の概要

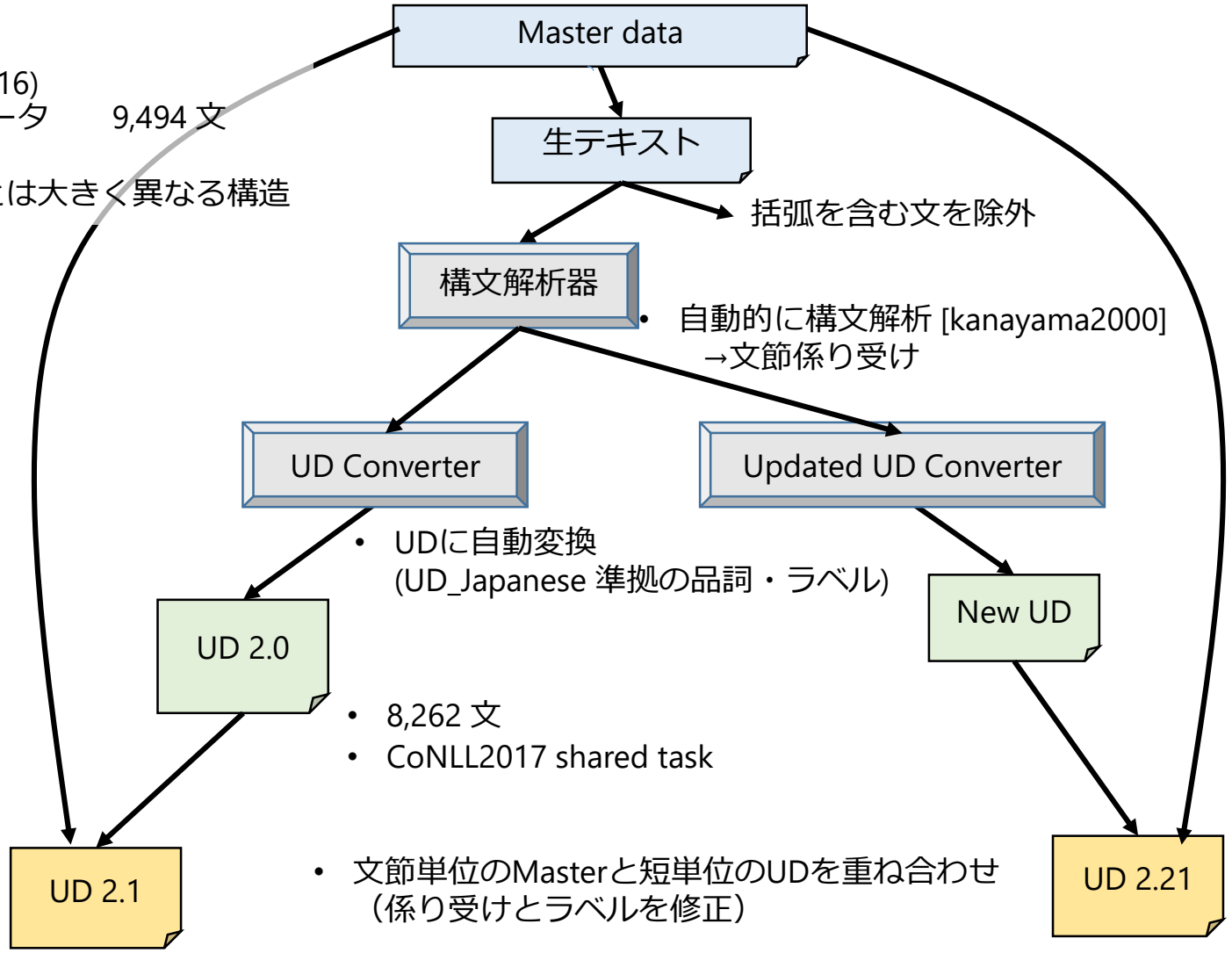
- 日本語UDコーパスの作成の経緯
  - UD Japanese_GSD を中心に
- コーパス作成の手順
  - 自動的な構文解析と元データの重ね合わせ
- 最近の改良について
  - （日本語UDチーム向け）
- 今後の課題

UDコーパスの歴史

- http://universaldependencies.org/
- V1.0 2014年11月
- V1.1 2015年5月
  - UD Japanese-KTC公開（毎日新聞ライセンスが必要）
- V1.2 2015年11月
- V1.3 2016年5月
- V1.4 2016年11月
  - UD Japanese（無印）： Google・DFKI によりmasterデータ作成（word = 文節）
- V2.0 2017年3月
  - UD Japanese 無印を短単位レベルに改変（時間の都合により、自動アノテーション）
  - UD Japanese 無印 → UD Shared Task #1のデータ + UD Japanese PUD
- V2.1 2017年11月
  - UD Japanese-GSD/PUD 更新　(無印はGSDに改名）
  - 自動アノテーションの結果を、文節レベルのmasterと重ね合わせ
- V2.2 2018年3月
  - → UD Shared Task #2のデータ
  - UD Japanese-BCCWJ, UD Japanese-Modern 公開
- 2018年6月現在
  - 修正版を作成中

Universal Dependencies Japanese-GSD

- Annotated by DFKI (Dec 2016)
- 主に日本語Wikipedia のデータ　　9,494 文
- 文節単位
  →日本語UDチームの議論とは大きく異なる構造

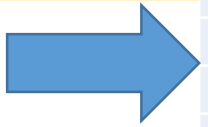


2018年5月の議論で
多くの問題点が判明

構文解析器→UDへの変換

text = 海は油膜を貼って青白く光っており、無数の漂着物が流れている。

[1] : 海(19) は(85)　(meishi 15/ kei_joshi 4 ＝) → [8]
[2] : 油膜(19) を(77) (meishi 2/ kaku_joshi 4 ＝) → [3]
[3] : 貼(9) っ(29) て(69)　　(doushi 5/ doushi 1 テ) → [5]
[4] : 青白(17) く(42) (keiyoushi 5/ keiyoushi 1 用) → [5]
[5] : 光(9) っ(29) て(69) お(9) り(28) 、(100)　　(doushi 5/ add_v 1 用) → [8]
[6] : 無数(19) の(76) (meishi 2/ joshi_no 6 ＝) → [7]
[7] : 漂着物(19) が(75)　　(meishi 2/ kaku_joshi 4 ＝) → [8]
[8] : 流れ(12) て(69) い(12) る(31) 。(100)　(doushi 5/ add_v 1 ＝) → [0]



1	海	海	NOUN	16	nsubj
2	は	は	ADP	1	case
3	油膜	油膜	NOUN	5	obj
4	を	を	ADP	3	case
5	貼っ	貼る	VERB	8	advcl
6	て	て	SCONJ	5	mark
7	青白く	青白い	ADJ	8	advcl
8	光っ	光る	VERB	16	advcl
9	て	て	SCONJ	8	mark
10	おり	おる	AUX	8	aux
11	、	、	PUNCT	8	punct
12	無数	無数	NOUN	14	nmod
13	の	の	ADP	12	case
14	漂着物	漂着物	NOUN	16	nsubj
15	が	が	ADP	14	case
16	流れ	流れる	VERB	0	root
17	て	て	SCONJ	16	mark
18	いる	いる	AUX	16	aux
19	。	。	PUNCT	16	punct

v2.0+master の重ね合わせ (→v2.1)

master
（文節単位）

# sent_id s26/ja				
1	海は	NOUN	NOUN-PRT	5 nmod
2	油膜を	NOUN	NOUN-PRT	3 nmod
3	貼って	VERB	VERB	0 root
4	青白く	ADJ	ADJ	5 advmod
5	光っており	VERB	VERB-AUX	3 conj
6	、	PUNCT	.	3 punct
7	無数の	NOUN	NOUN-POSTP	8 nmod
8	漂着物が	NOUN	NOUN-PRT	9 nmod
9	流れている。	VERB	VERB-AUX-AUX	3 conj

左主辞の並列構造

Python スクリプトにより変換

v2

# sent_id = dev-s23				
1	海	海	NOUN	16 nsubj
2	は	は	ADP	1 case
3	油膜	油膜	NOUN	5 obj
4	を	を	ADP	3 case
5	貼っ	貼る	VERB	8 acl
6	て	て	SCONJ	5 mark
7	青白く	青白い	ADJ	8 acl
8	光っ	光る	VERB	16 acl
9	て	て	SCONJ	8 Mark
10	おり	おる	AUX	8 aux
11	、	、	PUNCT	8 punct
12	無数	無数	NOUN	14 nmod
13	の	の	ADP	12 case
14	漂着物	漂着物	NOUN	16 nsubj
15	が	が	ADP	14 case
16	流れ	流れる	VERB	0 root
17	て	て	SCONJ	16 mark
18	いる	いる	AUX	16 aux
19	。	。	PUNCT	16 punct

v2.1

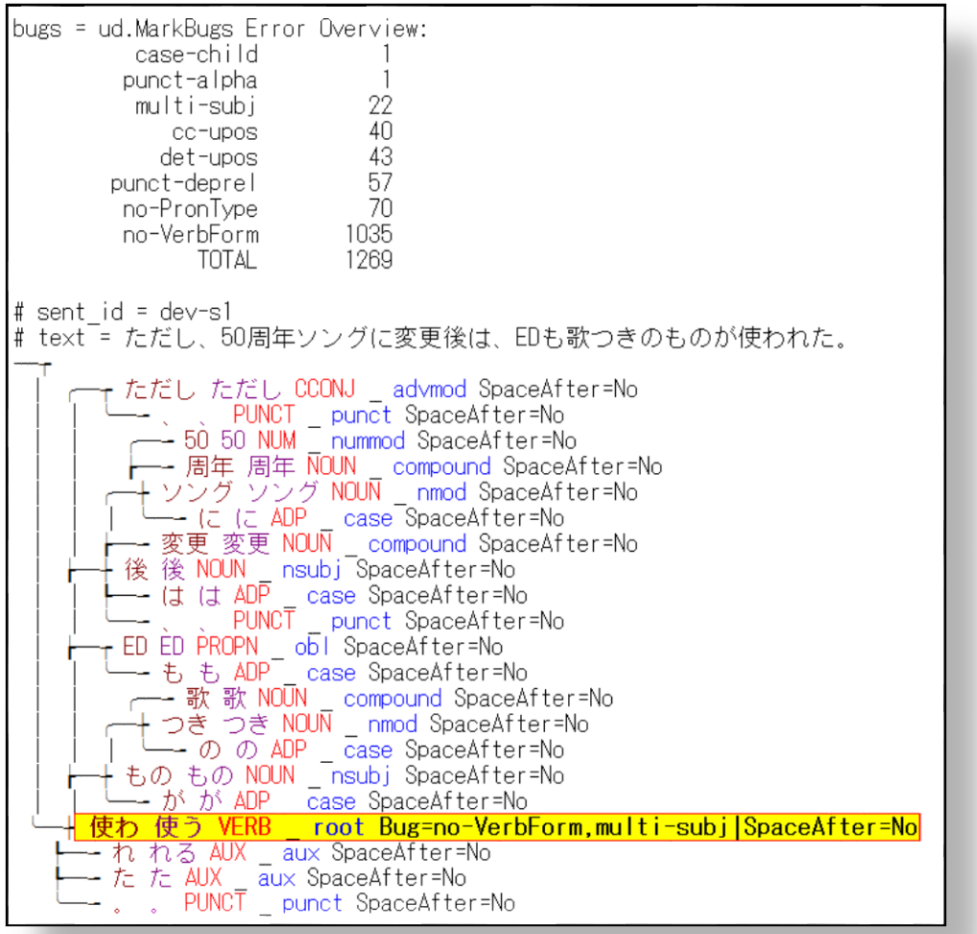
# sent_id = dev-s23				
1	海	海	NOUN	8 nsubj
2	は	は	ADP	1 case
3	油膜	油膜	NOUN	5 obj
4	を	を	ADP	3 case
5	貼っ	貼る	VERB	8 advcl
6	て	て	SCONJ	5 mark
7	青白く	青白い	ADJ	8 advmod
8	光っ	光る	VERB	16 advcl
9	て	て	SCONJ	8 mark
10	おり	おる	AUX	8 aux
11	、	、	PUNCT	8 punct
12	無数	無数	NOUN	14 nmod
13	の	の	ADP	12 case
14	漂着物	漂着物	NOUN	16 nsubj
15	が	が	ADP	14 case
16	流れ	流れる	VERB	0 root
17	て	て	SCONJ	16 mark
18	いる	いる	AUX	16 aux
19	。	。	PUNCT	16 punct

advcl となるべき

text = 海は油膜を貼って青白く光っており、無数の漂着物が流れている。

udapiによるvalidation

- https://github.com/udapi/
- Reduced error cases
  - Circulation
  - Functional word chain
- Remaining errors
  - case-child 1
  - punct-alpha 1
  - multi-subj 22
  - cc-upos 40
  - det-upos 43
  - punct-deprel 57
  - no-PronType 70
  - no-VerbForm 1035



v2.21で解消した問題点 (2018年6月)

- 用言→体言のラベルが全て acl となっていたのを、acl / advcl に振り分け
- 連体詞「この」「その」などの品詞を ADJ→DETに修正
- 準体言「の」「ん」の品詞を PART (aux) →SCONJ (mark) に修正
- 文末体言止めに係る格要素を nmod → nsubj/obj/iobj/obl に修正
- 一部の句読点のラベルを compound→punctに修正
- 「Vとみられる」「Vと認める」などに ccomp を割り当て
- 長単位に基づく用言・体言の判定
  - 「大きさが　わかる」（ADJ←VERB）： advcl から nsubjへ
- 数字の修飾の nummod を修正

Known issue （細かい点）

- 「Vこと」が文節に含まれることによる異様な構造
  - 「食べるのが好き」「食べることが好き」
    「あの人のことが好き」「今日のことは秘密」
- 表層しか見ていない格
  - 「1995年に」などがiobjになっている
- 固有名詞の接続に flat が使われていない
- 半角・全角文字の表層の不整合
- 重複した文の削除（国外より指摘あり）

今後の課題

- Goldの形態素との重ね合わせ
  - 国立国語研での作業の結果
- XPOSの付与
- Featureの付与
- 並列の問題の解決
  - UDW で発表？

日本語チームでやるべきこと

- Quickな作業
  - Shared task などに間に合わせる
  - 2018年3月のリリースでupdateできなかったのを猛省
- 世界に向けた発信
  - 並列構造の問題など、UD韓国語チームと声を上げる
- UDを正しく使う・広げる努力
  - 日本語のデータは完成度が低い
    - 英語・ドイツ語・韓国語など他の言語もそんなもん
  - 学術界で使うには
  - 産業界で使うには