Universal Dependency に基づく多言語処理の共通化

金山 博
hkana@jp.ibm.com
日本アイ・ビー・エム株式会社 東京基礎研究所

1 はじめに

Universal Dependencies (UD) は、多言語で一貫性を持つツリーバンク（構文構造付きコーパス）を設計・提供するプロジェクトであり、多言語の構文解析器の実装、他の言語の資源を用いた言語横断的な学習や、言語間の定量的な比較の研究などに役立てることを目指している [11, 10]。2016 年 1 月現在では、33 言語のツリーバンク、23 言語の構文構造の定義が公開されている [12]。

近年の UD に関する報告におけるデータと処理の概念を図 1 に示す。まず、各言語の UD 準拠のツリーバンクを作成するために、既存のツリーバンクを変換する試み（ロシア語 [8]・イタリア語 [2]・スウェーデン語 [1] など）がある。日本語についても、既存の句構造コーパス [17] から変換したコーパスを公開している。また、特に資源が少ない言語を対象に、他の言語のツリーバンクを補完的に用いて品詞タグ付け [15] や依存構造解析 [5, 14] をするといった言語横断的な転移学習の手法が考案されている。これらはいずれも品詞タグ付けや構文解析の結果を UD のツリーバンクと比較することにより性能を評価している。

一方で、UD に基づいた構文解析の結果を、構文解析そのものの評価のためではなく、多言語を扱うアプリケーションにおいて活用したといった報告は、知る限り存在しない。また、日本語については、品詞や依存構造のラベルの定義に関する議論 [16] の中でも様々な問題点が残っており、UD 上で学習や評価をした日本語の解析器が実用に耐えるかどうかはまだ検証されていない。また、UD のツリーバンク上での精度評価だけが議論されると、アプリケーションにおいて有用な解析器の実現と乖離していくことも懸念される。

そこで本稿では、解析器そのものの作成や評価とは別の視点で、多言語間で共通の構文構造を、その後段の処理に活かせるかどうかについて考える。図 2 の概念図に示すように、複数の言語の構文解析の結果を UD に変換し、多言語共通の処理をした結果を、アプリケーション上で評価することにより、UD を介することの効果を検証する。

ケーススタディとして、英語やスペイン語向けに設計された質問応答システムにおいて、日本語の構文構造をどのように表現すれば後段の処理を再実装せずとも日本語の質問応答を動作されられるかを調査した。

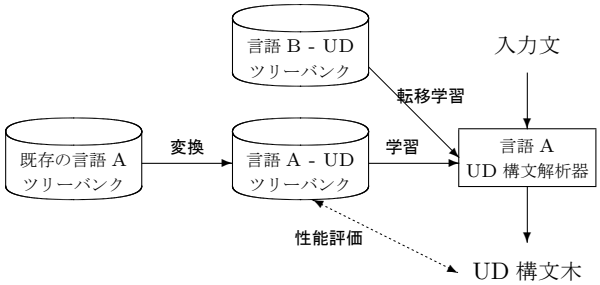


図 1　これまでの UD に関する研究の典型的な例。

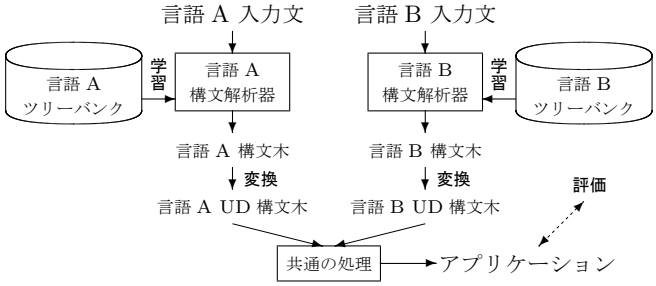
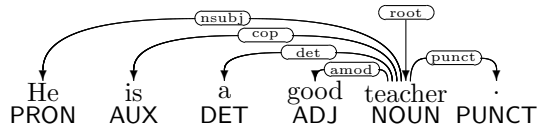


図 2　本稿で議論する、UD の構文木をアプリケーションの多言語化に活用するイメージ。

2 Universal Dependencies の特徴

UD では、英語の例 (1) のように構文構造を表す。文の主辞 (root) 以外の語は文中のいずれかの語に依存する形とするため、文全体は交差を許す木構造となる。



単語相互の依存関係だけを記述し、句構造 (constituent) を考慮しないことにより、表現が単純化され、言語資源の作成を省力化できるだけでなく、くだけた表現や特殊な言い回しに関して頑健な表現となるといった利点があると思われる。

全言語の品詞体系を集約するために Google Univer-

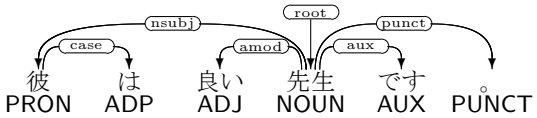
表 1　Universal PoS 2.0 の 17 種の品詞タグセット。＊は内容語とみなされるもの。

NOUN ＊	名詞
PROPN ＊	固有名詞
VERB ＊	動詞
ADJ ＊	形容詞
ADV ＊	副詞
INTJ	間投詞
PUNCT	句読点
SYM	記号
X ＊	その他

PRON	代名詞
NUM ＊	数詞
AUX	助動詞
CONJ	接続詞
SCONJ	従属接続詞
DET	限定詞
ADP	接置詞
PART	接辞

sal Part-of-speech Tags [13] を基にした 17 種の品詞セット（表 1）を用いる。また、依存構造のラベルとして Universal Stanford Dependencies [4] で定義された 42 種のラベルを一部改変した 40 種を用いる。

主語と動詞の呼応などの文法的な対応関係を扱わずに、内容語相互の関係を重視することにより、言語間の構造の違いを吸収している。その典型的な例が、例文 (1) にもあるコピュラの扱いである。be 動詞を主辞、he と teacher はそれぞれ be の主語と補語だと捉える従来の文法と異なり、UD では teacher を文の主辞 (root) とし、he と teacher の間に直接の依存関係を付与している。これにより、欧米語特有の be 動詞に特化しない依存構造を得ることができる。対応する日本語の UD の構造 (2) と比較すると、構文構造の形や品詞は異なれど、文の主辞が「先生」であることや、「彼 ← nsubj −先生」（名詞の主語）、「良い← amod −先生」（形容詞の修飾）といった内容語相互の関係が両言語で共通となっていることがわかる。

(2)


3　UD の問題点

2 節の通り、UD の構造は、異なる言語について一定の類似性を持たせて表現できるが、日本語の UD の設計 [16] の際には多くの課題に直面した。具体例として以下のようなものがある。

(A) 格の扱い

UD には英語の主格・目的格・間接目的格に対応して nsubj, dobj, iobj といったラベルがあるが、助詞などで格機能を表す日本語においてはこれらのラベルを区別して付与することは自明ではない。

(B) 態の扱い

UD では受動態に関連した特別なラベル (nsubjpass, auxpass など) を与えているが、日本語では受動態だけを特別視する合理的な理由が無い。格の交替に関する現象を捉えるなら使役なども区別する必要がある。

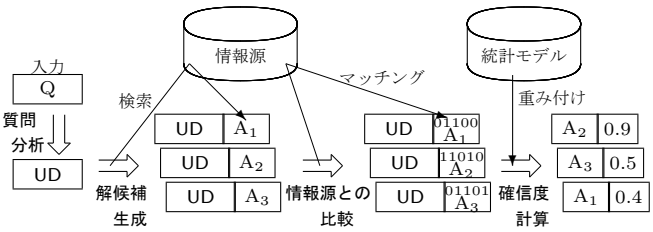


図 3　多言語質問応答の処理の流れ。

(C) 用言の連体形の表現

英語における限定用法の形容詞は amod で表されるが、日本語の形容詞の連体形は、節による連体修飾 acl との明確な線引きが難しい。

(D) 品詞の不一致

same, different(いずれも品詞は ADJ) に対応する日本語の語が「同じ」(NOUN)・「違う」(VERB) となるように、意味的な対応が取れる現象についても品詞が異なる。また、「製品を 発表。」といったサ変名詞の体言止めを NOUN、VERB のどちらの品詞とみなすかなども判断が難しい。

(A) の例は、他の言語（特に英語）の構文構造に即したラベル付けにより、リソース作成の自動化や一貫性の保持に影響を及ぼす原因となる。(B)(C) は、言語ごとに記述すべき事象の詳細さの度合いが異なるという点である。特定のラベルを一切用いないことは容易だが、言語横断的な学習をする際には障害となろう。決して interlingua のような共通化を目指すわけではないとはいえ、(C) や (D) のように、近い意味を持つ対訳であってもラベルが異なることが避けられないことには注意を要する。従って情報抽出のルール等の共通化には限界がある。

このような制限が残っている中で、いかに UD を活用できるかを、4, 5 節で考える。

4　質問応答の多言語化

UD を活用する事例として、質問応答システムの多言語化を考える。オープンドメインの質問応答システム Watson[6] では、英語の構文解析器 ESG [9] が多大な貢献をしていた。一方で、基本となる構文構造が英語に特化していると、その出力を受け付ける多くのコンポーネントを多言語化する際に障害となるため、UD に基づく構文木をもとに後段の処理を動かす試みがなされている [3]。

ここで扱う質問応答の流れを図 3 に示す。その多言語化にあたって、言語に依存する処理を最初の質問分析のプロセスに集約し、その後段のコンポーネントを共通化する。多言語向けに簡略化された手順は以下のようになる。

質問の分析　入力された質問文を構文解析したものを、

UD の構造で表現しておく（図 3 で UD と示す）。

解候補の生成 質問文に含まれる語句をクエリとして情報源から文書検索をして、記事のタイトルや頻出するリンク先などを解の候補とする。検索クエリは UD の構文木から内容語（表 1 参照）を取り出して生成する。情報源として Wikipedia を用いれば、各言語で構造が共通であるため、解の候補の生成のロジックの共通化が容易である。

情報源との比較 情報源から検索して得たパッセージと質問文との間の一致の度合いを数値化して、それぞれの解候補の素性とする。その際に、UD の品詞や依存構造とそのラベルが参照される。

確信度の計算 質問と正答のペアからなる学習データを用いた機械学習により、上記の処理で出力される素性に対して重み付けを行う。これにより各候補の確信度を求め、確信度が最大の候補がシステムの出力となる。このプロセスはもとより言語非依存である。

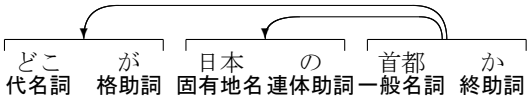
このアプローチにより、スペイン語やポルトガル語の質問応答に対しては英語のコンポーネントを再利用できていたが、構文構造や語の単位が大きく異なる日本語についても同様にシステムが簡単に構築できるか、また UD の性質との関連について、5 節で考察する。

5 日本語 UD の活用と効果

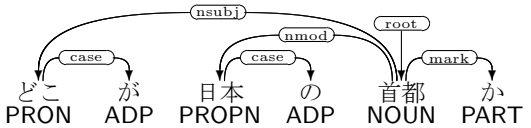
5.1 構文解析結果の変換

日本語の UD の構造を得るために、既存の日本語構文解析器 [7] の文節係り受けの出力を、日本語 UD の定義 [16] に即した変換を施した。但し、単語の単位については変換元の形態素解析の結果から変更していない。例として、文節係り受け (3) は UD の依存構造 (4) に変換される。

(3)



(4)



UD で表現した構文構造を、図 3 の質問分析の出力として用いて、その後は他の言語と共通の処理を行うことにより質問応答の結果を得る。

5.2 実験

UD 準拠の構文木を多言語で共通の処理の入力として用いることの効果を見るために、構文解析の出力を改変して質問応答のパイプライン全体を動作させて、質問応答の正解率の変化を観察した。評価と学習には、英語のオープンドメインの質問セットを日本語に翻訳したもの、情報源には日本語 Wikipedia を用いる。なお、日本語に特化した処理で性能を上げる余地は多いが、それについては今回の対象外とする。

表 2　評価データ 220 問の質問応答の性能。

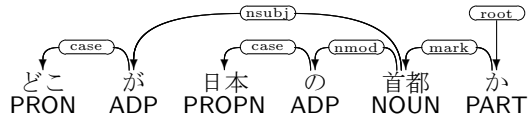
	Recall	Accuracy
UD 準拠	67.3%	15.0%
従来型依存構造	62.7%	10.5%
依存構造ラベル無し	62.7%	10.0%
品詞タグ無し	58.6%	14.5%
ランダム品詞	34.1%	2.7%
検索のみ	41.3%	3.6%

構文解析の出力として以下のようなものを用意した。当然ながら、データ構造自体はシステムが想定するものでないと動作させられないため、形式は保ちつつ、ラベルや依存構造を改変した状況を人為的に作成している。

UD 準拠　上記の例 (4) のように、UD の定義に従って変換した構造。

従来型依存構造　依存構造を (5) のような従来型の日本語単語係り受けの構造に置き換え、それに応じてやや怪しげなラベルを振る (本来の UD の制約とは異なる)。

(5)



依存構造ラベル無し　依存構造のラベルを、全てデフォルト値である dep に置き換える。

品詞タグ無し　品詞のタグを、全てデフォルト値である X に置き換える。

ランダム品詞　品詞タグをランダムに付与する。

表 2 に、正解が解候補の中に現れた割合 (Recall) と、正解に対して最も高い確信度が付与された割合 (Accuracy) を示す。ベースラインとしての「検索のみ」は、質問文から内容語を抽出したクエリで Wikipedia の記事を検索し、そのタイトルを解候補とした場合の正解率を示す。この値から、自明に解ける問題は少ないことがわかる。

質問応答の精度自体は決して高くはないが、構文解析以外の日本語対応を行うことなく全体のシステムを動作させることができた。依存構造やラベルが想定するものと異なると、クエリ作成時の語の選択や重み付けが効かなくなることにより Recall が低下し、２つの内容語の間の依存構造の一致による類似度の判定がしづらくなる[*1]ことにより Accuracy が低下した。

品詞をすべて X にした場合、全単語が内容語とみなされるため、検索結果にノイズが出て Recall が下がるが、構文構造が正しければ正解に確信度を与えられるため、Accuracy の低下は最小限だった。品詞をランダ

[*1] 例えば (4) では得られる「日本← nmod －首都」の依存関係が (5) の構造では得られず、マッチングに使うことができない。

ムにした場合は、検索のクエリに必須となる内容語が同定できなくなるために、性能が大きく低下した。

6 まとめと今後の課題

本稿では、Universal Dependencies の出力が、多言語を扱うアプリケーションの処理の共通化にどのように寄与するかを調べた。まず、構文解析の結果を UD に変換するだけでも後段の処理がそのまま適用できることを確認し、また他の言語向けに設計されたシステムが想定する通りに UD で定義した構文構造であったほうが end-to-end のシステムで高い性能が得られることがわかった。

今回は後段の処理が簡素化されており、品詞や依存構造のラベルを使う局面は、検索クエリの作成や、質問文と情報源の比較などに限られ、特定のラベルを用いて辞書引きや推論をしたりはしていない。深い処理が使われるようになると、格構造解析に相当する nsubj, dobj, iobj の区別ができるか、否定の表現を neg ラベルで捉えられるかといった正確さが重要になると考えられる。

また、今回は UD の定義のみを用いており、UD のツリーバンクは利用していない。UD の構文解析の精度を評価し、UD を利用したアプリケーションの性能との相関も調べてみたい。

このように複数の言語で評価ができるアプリケーションが増えていくと、多言語の構造を設計するにあたっての方針の策定、すなわち言語表現の多言語化の妥当さを定量的に測定できるようになると期待される。

参考文献

[1] Lars Ahrenberg. Converting an English-Swedish parallel treebank to Universal Dependencies. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 10–19, 2015.

[2] Cristina Bosco, Simonetta Montemagni, and Maria Simi. Converting Italian treebanks: Towards an Italian Stanford dependency treebank. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 61–69, 2013.

[3] Keith Cortis, Urvesh Bhowan, Ronan Mac an tSaoir, D.J. McCloskey, Mikhail Sogrin, and Ross Cadogan. What or who is multilingual Watson? In *Proceedings of COLING 2014: System Demonstrations*, pp. 95–99, 2014.

[4] Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of LREC*, pp. 4585–4592, 2014.

[5] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 845–850, 2015.

[6] D. A. Ferrucci. Introduction to "This is Watson". *IBM Journal of Research and Development*, Vol. 56, No. 3.4, pp. 1:1–1:15, 2012.

[7] Hiroshi Kanayama, Kentaro Torisawa, Yutaka Mitsuishi, and Jun'ichi Tsujii. A hybrid Japanese parser with hand-crafted grammar and statistics. In *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 411–417, 2000.

[8] Janna Lipenkova and Milan Souček. Converting Russian dependency treebank to Stanford typed dependencies representation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics(EACL)*, pp. 143–147, 2014.

[9] Michael C. McCord, J. William Murdock, and Branimir K. Boguraev. Deep parsing in Watson. *IBM Journal of Research and Development*, Vol. 56, No. 3.4, pp. 3:1–3:15, 2012.

[10] Ryan T McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith B Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. Universal Dependency annotation for multilingual parsing. In *ACL (2)*, pp. 92–97, 2013.

[11] Joakim Nivre. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*, pp. 3–16. Springer, 2015.

[12] Joakim Nivre, Cristina Bosco, Jinho Choi, Marie-Catherine de Marneffe, Timothy Dozat, Richard Farkas, Jennifer Foster, Filip Ginter, Yoav Goldberg, Jan Haji, Jenna Kanerva, Veronika Laippala, Alessandro Lenci, Teresa Lynn, Christopher Manning, Ryan McDonald, Anna Missila, Simonetta Montemagni, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Maria Simi, Aaron Smith, Reut Tsarfaty, Veronika Vincze, and Daniel Zeman. Universal dependencies 1.0, 2015.

[13] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of LREC*, 2012.

[14] Jörg Tiedemann. Cross-lingual dependency parsing with universal dependencies and predicted PoS labels. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pp. 340–349, 2015.

[15] Guillaume Wisniewski, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[16] 金山博, 宮尾祐介, 田中貴秋, 森信介, 浅原正幸, 植松すみれ. 日本語 Universal Dependencies の試案. 言語処理学会第 21 回年次大会発表論文集, pp. 505–508, 2015.

[17] 田中貴秋, 永田昌明, 松崎拓也, 宮尾祐介, 植松すみれ. 統語情報と意味情報を統合した日本語句構造ツリーバンクの構築. 第 20 回言語処理学会年次大会発表論文集, pp. 737–740, 2014.