

多言語評価表現抽出を通じた Universal Dependencies の検証

金山 博

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

岩本 蘭

慶應義塾大学理工学研究科
r.iwamoto@keio.jp

1 はじめに

Universal Dependencies (UD)[6] 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、2019年11月現在、90言語・157のツリーバンクが公開されている。すべての言語において、構文構造は17個の品詞タグと37種の係り受けラベルを付与した依存構造木で表現され、そこに言語固有の素性が加えられている。UDを用いた構文解析のシェアドタスク [11] の結果、多言語の構文解析の技術が大きく進歩したほか、共通の構文構造とBERTを用いた全言語単一のモデルによる構文解析も高い性能を示している [5]。さらにUDは言語学的な比較にも活用されている [1, 2]。

一方で、UDに基づく構文解析の結果を多言語処理に応用するという観点で議論した研究は僅かしか見当たらない [4]。そこで本稿では、評価表現抽出を題材として、UDの効用や問題点について議論する。文中にある評価表現を正確に抽出するにあたって、既存の統計的手法で解決されない文法構造があること [9] や、構文構造と文法規則の活用には価値があること [10] が示されており、評価表現抽出は構文解析の一つの理想的な応用事例といえる。2節で構文構造を用いた評価表現抽出の手法、3節でその多言語化について述べる。4節の実験では、UDの構造によって共通化される処理と言語依存の処理の割合や、UDのコーパスの更新がアプリケーションにもたらす影響について観察、議論する。

2 評価表現抽出

本稿で抽出したい評価表現とは、文中にある好評ないし不評を示す節であり、述語と対象を要素とする。英語の例文 (1) は、hotelに関する好評の表現 (1a) と waiterに関する不評の表現 (1b) を含んでいる。

- (1) I love the hotel but she said none of the waiters were friendly.
 (1a) + love (hotel)
 (1b) - not friendly (waiter)

これらの評価表現を検出するために、図1に示すUDの木構造を辿りながら評価表現を構築していく。これは、構文木の根からトップダウンにパターンと辞書を適用する手法 [13] をベースとしている。

2.1 節の検出

まず、文の主節の主辞である root のノードの語と項を、2.2節で後述する手法により辞書引きして、評価の

	lemma	品詞	極性	格フレーム
(a)	love	VERB	+	nsubj, obj
(b)	friendly	ADJ	+	nsubj
(c)	unhappy	ADJ	-	nsubj, with
(d)	high	ADJ	-	nsubj:“price”
(e)	noise	NOUN	-	

表1 極性辞書の語彙項目の例。‘+’は好評、‘-’は不評の極性を表す。

極性および対象を同定する。また、一つの文の中にある複数の評価表現を検出するために、並列構造を示す conj, parataxis などのラベルで係る子ノードを再帰的に調べる。“say” や “think” のように ccomp を子に持つ動詞であった場合、その子ノードを調べる。これらの操作により文 (1) から2つの評価表現 (“love” 及び “friendly” の否定) が検出される。

conjに加えて、“because”、“though”といった接続詞を mark のラベルで従える節が advcl ラベルで係っている場合もカバーすることにより、文 (2) から2つの評価表現を検出する。ccompと同様に、文 (3) にある “make” の xcomp の子ノードも探索の対象とする。^{*1}

- (2) [+] I was **satisfied** because amenities were **great**.
 (3) [+] He made the travel **excellent**.

2.2 極性辞書とのマッチング

先に述べた方法により着目するノードを定め、表1に示すような、lemma・品詞と格フレームの組に対して極性が割り当てられた語彙項目を用いて、節の極性・述語と対象を同定する。(a)は主語 (nsubj) と直接目的語 (obj) を取る動詞 “love” で、好評の極性を持ち、評価の対象は直接目的格の語である。(b)は形容詞 “friendly” が好評であること、(c)は “unhappy” が不評の極性で、前置詞 “with” を伴う句に対象があることを表す。(d)のように、“high”は主語が “price” の時にのみ不評の極性を持つことも定義できる。なお、これらの形容詞の語彙項目は、“high price”のように amod で修飾する名詞句とも符合する。(e)は名詞の項目で、名詞述語節の主辞や “increase” の直接目的語 (obj) となった時に不評の極性を与える。

2.3 否定による極性の反転

極性を反転させる否定の表現としてさまざまなタイプが存在する。最も基本的なものは文 (4) のように advmod ラベルを用いた述語の否定であり、“not”のほか、“never”、さらに “seldom” “scarcely”などを列挙

*1 太字・下線はそれぞれ評価の述語と対象を示す。

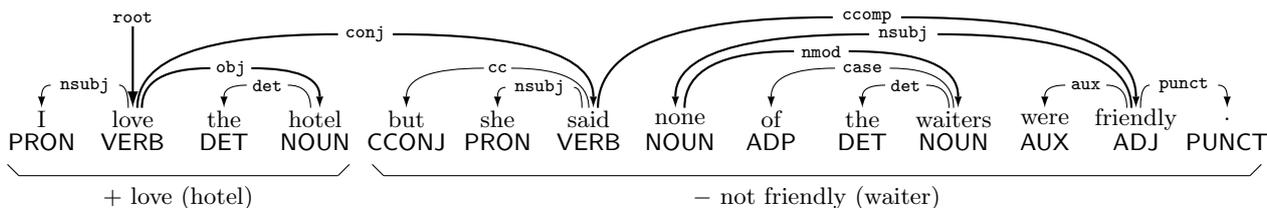


図1 文(1)に対応する依存構造木。root ノードから太い矢印で示す依存構造をたよりに木構造を辿りながら、2つの評価表現の極性・述語・対象の組を検出している。

ar	PADT	fr	GSD	ko	GSD
cs	PDT	he	HTB	nl	Alpino
de	GSD	id	GSD	pt	Bosque
en	EWT	it	ISDT	tr	IMST
es	Ancora	ja	GSD	ru	SynTagRus
				zh	GSDSimp

表2 本研究で用いる16言語のUDコーパス。

してある。文(5)は名詞の否定で、“no”などがdetラベルで名詞に係る時に極性が反転される。

- (4) [-] The hotel was **not good**.
 (5) [+] It was **no problem**.

さらに、(1)で見たように、“none of...”や“nothing”が主語や目的語となった時にも、否定とみなして述語の極性を反転させる。

3 システムの多言語化

2節で見た、英語向けの評価表現抽出器を、UDの構文構造を活用して他の言語に対応させていく。

3.1 UDコーパスの活用

今回は表2に示す16言語のUDコーパスで学習された構文解析器を用いる。中国語のGSDは繁体字がベースとなっているので、簡体字を扱うためにGSDSimpを利用した。

UD_Arabic-PADTのコーパスでは、lemmaに発音記号が付与されるため、辞書と符合される際にはこれを外す。韓国語UDでは、単語の単位が語節になっており、lemmaは語節内の形態素を‘+’マークで結合するという特殊な形になっている。この形では語彙項目との符合が困難なため、lemmaの文字列を‘+’で分割した最初の単位に語尾“da”を付加して辞書と比較する。

3.2 辞書の作成

2.2節で見た英語の極性辞書を、機械翻訳や多言語の分散表現を用いて他の言語に変換した。この手法については別項[14]を参照されたい。

3.3 構文木上の操作

次に、UDの構文木上の操作を他の言語に適用させる。本稿では、部分木からの節の検出と、否定による極性の反転に関する操作を中心に、言語間で共通のもの、言語毎に語彙を列挙するもの、各言語に特化したものに分類しながら解説する。

節の検出 評価表現を同定する節を探索する操作の基本的な部分は、UDの構造により言語間で共通となる。まず、文中でrootとなる語を起点に、並列構造を示す

conj、parataxis及びlistのラベルに係る子ノードを再帰的に探索する。

節補語など(ccompとxcomp)を取る動詞は当然ながら言語ごとに異なる。英語なら“think”、“make”、スペイン語では“creer”(‘believe’)や“parecer”(‘seem’)、日本語では“思う”、“感じる”などの語彙を列挙した。

日本語UDでは、UDの原則である左側主辞の並列構造が日本語の構文構造に合わないため、conjラベルを用いていない[3]。そこで、文中の複数の節を調べるために、advclラベルにより修飾する従属節を走査する。**極性の反転** 否定の表現を扱う言語共通の方法として、素性の中にある‘Polarity=Neg’の利用が考えられる。これは、今回対象とする言語の中では、ar, cs, de, es, fr, he, id, pt, tr, zhの10言語で使われており、ドイツ語の“nicht”や中国語の“不”などの否定の副詞のほか、アラビア語の否定動詞、チェコ語・トルコ語の述語の否定形などに付与されており、言語毎の語彙を列挙することなく否定の表現を捕捉できる。

11言語(de, en, es, fr, he, id, it, nl, pt, ru, zh)では、英語の“not”のようにadvmodで述語を修飾する語や、“no”のようにdetで名詞を修飾する語を列挙することにより否定を検出できる。

その他の言語では否定の表現の方法が異なり、個別の対応が必要である。日本語の“ない”やフィンランド語の“ei”は、否定の助動詞で、auxラベルで述語に係る。韓国語では、用言と否定語の組合せが複合語と見なされて、その間はflatの依存関係となっている。

4 実験と観察

機械学習の手法と異なり、辞書や文法規則に基づく手法は、性能の定量的な評価や他の手法との比較が難しく、また評価データの質の影響を受けやすい。本稿では、言語間で共通する簡単な構造の評価データを用いて、言語毎の構文構造上の操作を定性的に評価し、UDによる共通化の度合いを観察していく。

4.1 実験設定

多言語の評価表現抽出の性能を測るため、16言語のaspectベースの評判分析の各タスクで使われてきたコーパスを、文単位の極性付きデータに変換した。詳細は別稿参照[14]。

すべての入力文には好評ないし不評のラベルが付与されているのに対し、我々のシステムは任意の個数の評価表現を検出する。適合率は、システムが出力した極性が正解の極性と同一である割合、再現率は、正解

言語	F ₂	部分木 (再現率)					否定 (適合率)				
		無	△ 共通	△ 語彙	△ 個別	全	無	△ 共通	△ 語彙	△ 個別	全
ar	72.4	21.2	+14.0	+1.6	0	36.8	94.2	0	0	+1.3	95.5
cs	68.8	25.0	+11.2	+0.4	0	36.6	77.6	+7.2	+3.0	+0.5	88.3
de	78.3	35.7	+9.4	+1.4	+0.3	46.8	92.4	+1.2	+0.6	0	94.2
en	77.5	32.8	+11.4	+2.0	+0.6	46.8	84.2	0	+8.0	+0.5	92.7
es	69.8	30.8	+5.4	+0.2	+0.2	36.6	80.8	+6.8	+2.1	+0.5	90.2
fr	74.2	31.8	+11.2	+0.6	0	43.6	85.3	-0.1	+0.5	+4.3	90.0
he	45.6	14.4	+1.8	+0.2	0	16.4	78.0	+4.0	0	0	82.0
id	58.7	29.4	+3.0	0	+1.4	33.8	88.8	+1.8	+2.1	0	92.7
it	53.4	23.6	+6.0	+0.2	0	29.8	81.0	0	+7.3	0	88.3
ja	68.0	32.1	0	0	+1.1	33.2	86.9	0	0	+5.3	92.2
ko	35.4	10.3	+0.6	0	0	10.9	79.1	0	0	+1.5	80.6
nl	73.4	30.4	+10.2	+1.0	0	41.6	84.0	0	+5.9	+0.9	90.8
pt	63.3	22.4	+6.6	+3.0	+0.4	32.4	75.0	+6.5	+0.6	+1.1	83.2
ru	64.7	19.6	+10.6	+0.2	0	30.4	81.1	0	+7.6	+1.4	90.1
tr	48.7	13.4	+2.6	0	+1.0	17.0	81.3	+3.3	+3.4	+3.2	91.2
zh	58.1	13.6	+9.6	+1.4	0	24.6	79.4	+5.0	0	+3.8	88.2
追加した操作		conj		ccomp	その他		Pol=Neg		advmod	その他	
		parataxis		xcomp					det		

表3 16言語の評価表現抽出の再現率と適合率。部分木からの節の検出・否定の扱いについて、対応が無い場合、言語によらず共通の方法・言語別に列挙した語彙・各言語の個別の処理を加えた時の改善の度合いとともに示す。「無」に対する3種の改善の累積が「全」の値であり、これが現状のシステムの性能である。

の極性と同一極性を持つ評価表現が検出された割合*2である。文中に複数の評価表現がある場合や、述語や対象の評価ができるわけではないが、極性を持つ節の抽出の評価としては充分であると判断した。

総合的な評価指標として、適合率を重視したF₂値(式(6)でβ=2としたもの)を用いる。これは、F₁値を用いた場合、文法構造を意識せずに好評・不評を含む単語を抽出することにより簡単に数値を上げられることと、評価表現抽出のユースケースでは適合率が特に重要である場合が多いことが理由である。

$$F_{\beta} = (1 + \beta^2) \frac{\text{prec} \cdot \text{rec}}{\text{prec} + \beta^2 \cdot \text{rec}} \quad (6)$$

4.2 言語別の評価と各操作の効用

16の言語の入力文をStanfordNLP[7]を用いて構文解析したのち、2節・3節で述べた手法により評価表現を抽出した時の性能を表3に示す。

まず、総合的な指標であるF₂値を見る。コーパスの分野や質と辞書の変換の影響があるため、言語間の直接的な比較ができるわけではないが、多くの言語において英語と同程度、ドイツ語はそれを超える値が得られており、語彙の変換や文法構造の多言語対応に概ね成功しているといえる。特筆すべきは韓国語の性能、特に再現率が低いことである。これは、韓国語のUDにおける語の単位が語節となっており、lemmaが特殊な形をしているために他の言語と同様の符合ができないことが原因であり、韓国語UDの表現方法の改善が待たれる。ヘブライ語はニュース分野の性質上、トルコ語は語彙の変換後の辞書のサイズが小さいことにより再現率およびF₂値が小さくなっている。

表3の「部分木」の項目は、文のroot以外から評価表現の節を抽出するための操作を外した場合と、言語

*2 本手法で評価表現を検出しない場合は常に偽となる。特に間投詞など、本手法で対象外の現象が数値に影響する。

間共通・語彙の列挙・個別の操作を加えることによる再現率の変化を示す。「△ 共通」は、conjなどのラベルを見て並列構造を認識することによる、「無」の場合との差分を示す。全言語共通の操作により、conjラベルが全く用いられない日本語以外の15言語において再現率が回復している。「△ 語彙」は、ccompとxcompラベルの節補語を取る語を列挙した場合で、12言語、特に英語とポルトガル語で効果があった。「△ 個別」はその他の言語別の操作であり、インドネシア語や日本語ではこの効用が大きい。

表3の「否定」の部分で、否定の現象を扱わない場合(無)からの適合率の変化を見よう。言語共通の素性として‘Polarity=Neg’がある。チェコ語・スペイン語・ポルトガル語・中国語ではこの素性で否定表現の過半が捉えられる一方、7言語では変化が無い。フランス語では典型的な否定形‘ne ... pas’の双方の語に否定の素性があり、二重否定と誤判定されることにより負の貢献となっている。アラビア語の否定の素性が出力されるのは機能語‘laa’に係る語にほぼ限られており、今回の規則では利用できなかった。advmodとdetに係る否定の語彙を列挙することにより、多くの言語の否定表現がカバーされるが、日本語・韓国語においては全ての否定表現が言語固有の形式となっている。

全体的に、構文的操作の多くは言語共通のものや、語彙の列挙により解決され、UDの構文構造の共通性が活かされている。しかし、特に日本語UD[12]は共通性が十分に活かされていない場合が多く、アノテーション基準のさらなる検討が必要であろう。

4.3 UDのバージョンの比較

UDのコーパスは3~6ヶ月ごとに更新されており、その影響が気になる。表4は、UDの3つのバージョンで学習されたUDPipe[8]のモデルで構文解析をした際の評価表現抽出の性能を示す。

日本語は最新のUD2.4が高い性能を示した。advcl

言語	UD2.0				UD2.2				UD2.4			
	LAS	適合率	再現率	F ₂	LAS	適合率	再現率	F ₂	LAS	適合率	再現率	F ₂
ar	64.3	88.4	15.8	46.1	65.1	85.3	16.8	47.0	66.6	83.7	17.0	46.9
cs	82.3	85.4	30.4	62.7	82.8	86.6	29.8	62.7	82.9	86.4	31.8	64.3
de	68.6	93.6	47.1	78.2	70.8	92.6	42.6	75.0	72.7	92.6	46.0	77.0
en	76.5	92.6	43.8	75.7	77.1	92.5	44.0	75.8	76.4	90.9	43.4	74.6
es	84.5	88.6	33.8	66.9	84.4	89.5	33.4	67.0	85.1	89.7	32.0	65.9
fr	80.7	88.7	39.4	70.9	81.0	90.5	39.8	72.1	84.5	91.3	39.8	72.5
he	57.9	82.1	14.0	41.6	57.9	82.1	14.0	41.6	58.3	82.3	13.2	40.2
id	74.3	-	0.0	-	74.4	93.2	30.4	66.0	74.5	92.4	31.6	66.7
it	86.1	85.7	25.6	58.3	86.3	89.1	25.0	58.9	86.7	85.5	25.2	57.8
ja	75.5	92.5	29.5	64.8	72.6	90.1	28.9	63.3	76.2	92.1	32.6	67.5
ko	60.5	-	0.0	-	61.4	83.3	9.1	31.7	61.4	83.7	8.2	29.5
nl	69.6	90.6	25.8	60.3	77.6	90.6	39.4	71.9	77.6	89.4	39.8	71.6
pt	82.5	81.9	29.2	60.2	82.2	79.9	27.0	57.4	82.7	78.7	27.8	57.6
ru	87.3	88.9	26.4	60.3	84.6	88.9	30.0	63.8	85.0	87.3	27.6	60.9
tr	55.8	92.6	15.2	45.9	54.0	94.7	14.6	45.2	55.1	94.8	14.8	45.6
zh	57.7	76.4	8.6	29.6	57.7	81.2	7.8	28.2	58.7	84.4	7.6	27.9

表4 UD コーパスの3つのバージョンにおける評価表現抽出の性能。UDPipe による構文解析の指標 (LAS; Labeled Attachment Score) とともに示す。zh は繁体字のコーパスで学習しているため再現率が低い。

と ac1 のラベルが正しく修正されたことなどが寄与している。オランダ語では、否定の副詞 “niet” の係り方やラベルが修正された UD2.2 で大きな改善が見られた。インドネシア語や韓国語の UD2.0 では lemma が割り当てられていなかったため、辞書との符合が全くできなかったが、UD2.2 以降で解決された。

更新に従って単調に性能が上がるのが期待されるものの、UD2.4 が最高の性能を示さない言語も多い。例えばロシア語では lemma の方針の変更により符合がしづらくなった。また、これら 16 言語において絶えず UD コーパスの更新が行われていることも確認された。

5 おわりに

本研究では、構文構造に基づく評価表現抽出を用いて、16 言語の UD コーパスの性質を観測した。こういった次工程での評価により、構文解析だけでは観測できない UD の質的な変化を知ることができ、これは今後のコーパスの見直しや、構文解析のあり方の再検討、新たな応用の設計などにも有用であると期待される。

参考文献

- [1] Aleksandrs Berdicevskis, Çağrı Çöltekin, Katharina Ehret, Kilu von Prince, Daniel Ross, Bill Thompson, Chunxiao Yan, Vera Demberg, Gary Lupyán, Taraka Rama, et al. Using universal dependencies in cross-linguistic complexity research. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 8–17, 2018.
- [2] William Croft, Dawn Nordquist, Katherine Looney, and Michael Regan. Linguistic typology meets universal dependencies. In *TLT*, pp. 63–75, 2017.
- [3] Hiroshi Kanayama, Na-Rae Han, Masayuki Asahara, Jena D Hwang, Yusuke Miyao, Jinho D Choi, and Yuji Matsumoto. Coordinate structures in Universal Dependencies for head-final languages. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pp. 75–84, 2018.
- [4] Hiroshi Kanayama and Koichi Takeda. Multilingualization of question answering using Universal Dependencies. In *Proceedings of Open Knowledge Base and Question Answering Workshop at SIGIR2017, Tokyo, Japan*, pp. 13–16, 2017.
- [5] Daniel Kondratyuk and Milan Straka. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of EMNLP-IJCNLP*, pp. 2779–2795, November 2019.
- [6] Joakim Nivre and et al. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC*, Portorož, Slovenia, 2016.
- [7] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 160–170, Brussels, Belgium, October 2018.
- [8] Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UD-Pipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99, Vancouver, Canada, August 2017.
- [9] Rohil Verma, Samuel Kim, and David Walter. Syntactical analysis of the weaknesses of sentiment analyzers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1122–1127, 2018.
- [10] David Vilares, Carlos Gómez-Rodríguez, and Miguel A Alonso. Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, Vol. 118, pp. 45–55, 2017.
- [11] Daniel Zeman, Filip Ginter, Jan Hajič, Joakim Nivre, Martin Popel, and Milan Straka. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Brussels, Belgium, 2018.
- [12] 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. Universal Dependencies 日本語コーパス. *自然言語処理*, Vol. 26, No. 1, pp. 3–36, 2019.
- [13] 金山博, 那須川哲哉, 渡辺日出雄. 木構造変換を利用した評判分析手法. *人工知能学会論文誌*, Vol. 26, No. 1, pp. 273–283, 2011.
- [14] 岩本蘭, 金山博. 多言語極性辞書の構築とその包括的評価. *言語処理学会第 26 回年次大会予稿集*, 2020.