Universal Dependencies の変遷と評価表現抽出への影響

岩本蘭
慶應義塾大学
r.iwamoto@keio.jp

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

1 はじめに

近年，自然言語処理の研究がいくつかの言語に偏っていることが問題視されており，多言語の言語資源作成やシステムの開発が進められている [1]．100 を超える言語のツリーバンクを提供する Universal Dependencies(UD) [2, 3] は言語間で共通の PoS タグや依存構造ラベルをアノテーションに用いている．UD の目的は言語間の共通性を保ちつつ個々の言語の性質を表現することであり，UD コーパスで学習した構文解析器は多言語の応用タスクへの活用が期待されている．UD のアノテーションは半年ごとに更新されているが，それらの更新が構文解析 [4] や応用タスクに与える影響については十分な分析がなされていない．

本稿では，23 言語の UD コーパス ver. 2.0 (2017 年 3 月) – ver. 2.7 (2020 年 11 月) を対象にし，コーパスの更新状況と傾向について分析する．次に UD のそれぞれのバージョンのコーパスで訓練した構文解析器 UDPipe [5] を用いて言及単位の評価表現抽出 [6, 7, 8] を行う．アノテーションの言語をまたがった一貫性がどの程度活用できているか，またコーパスの更新が良い方向に作用しているかどうかを調査することが目的である．

2 UD コーパスの更新

UD コーパスの内容の例を表 1 に示す．本来 10 項目からなるアノテーションのうち評価表現抽出の際に用いる項目のみを抜き出している．その中から ID を除く 6 項目について，表 2 に示す 23 言語の UD コーパスにおける更新とその影響を分析する．

コーパスの更新の割合を項目別，バージョン別に分けて図 1 に示す．比較する 2 つのバージョンの両方に存在する文を抽出し，重複を削除した後，語の各項目に違いがあった割合を色の濃さで表している．UD2.0–2.4 [1)] ではほぼ全ての言語で活発に更新

1） 以降，UD ver 2.x を UD2.x と示す．

表 1 UD コーパス (英語) の例．feats の出力は一部省略．

text = He eats apples.

ID	form	lemma	UPOS	feats	head	deprel
1	He	he	PRON	Case=Nom	2	nsubj
2	eats	eat	VERB	Mood=Ind	0	root
3	apples	apple	NOUN	Number=Plur	2	obj
4	.	.	PUNCT	_	2	punct

され，品詞タグや依存関係ラベルの変更が多い．それに対し UD2.4–2.6 ではいくつかの言語内でのアノテーション方針の大幅な変更や，より詳細な属性を記述する項目である feats の変更が顕著である．

UD のアノテーション方針に沿って言語間での一貫性を保つことと，その言語の特徴を正確に記述することのバランスは難しいが，それぞれの言語で改良を重ねられている．コピュラや助動詞に関するタグ付けを一例としてとりあげると，英語の UD2.5 では “have been” や “will be” の “be” は VERB から AUX へ変更され，オランダ語 (UD2.4) でも “hebben”(‘have’) に対し同様の変更が見られた．ポルトガル語 (UD2.5) では AUX タグをつけられた多く

表 2 分析した treebank と UD2.6 での文数

言語	コーパス名	文数
アラビア語 (ar)	PADT	7,664
カタルーニャ語 (ca)	AnCora	16,678
チェコ語 (cs)	PDT	87,913
ドイツ語 (de)	GSD	15,590
英語 (en)	EWT	16,622
スペイン語 (es)	Ancora	17,680
ペルシャ語 (fa)	Seraji	5,997
フランス語 (fr)	GSD	16,341
ヘブライ語 (he)	HTB	6,216
ヒンディー語 (hi)	HDTB	16,647
クロアチア語 (hr)	SET	9,010
インドネシア語 (id)	GSD	5,593
イタリア語 (it)	ISDT	14,167
日本語 (ja)	GSD	8,071
韓国語 (ko)	GSD	6,339
オランダ語 (nl)	Alpino	13,578
ノルウェー語 (no)	Bokmaal	20,044
ポーランド語 (pl)	LFG	16,746
ポルトガル語 (pt)	Bosque	9,364
ロシア語 (ru)	SynTagRus	61,889
スウェーデン語 (sv)	Talbanken	6,026
トルコ語 (tr)	IMST	5,635
中国語 (zh)	GSD	4,997

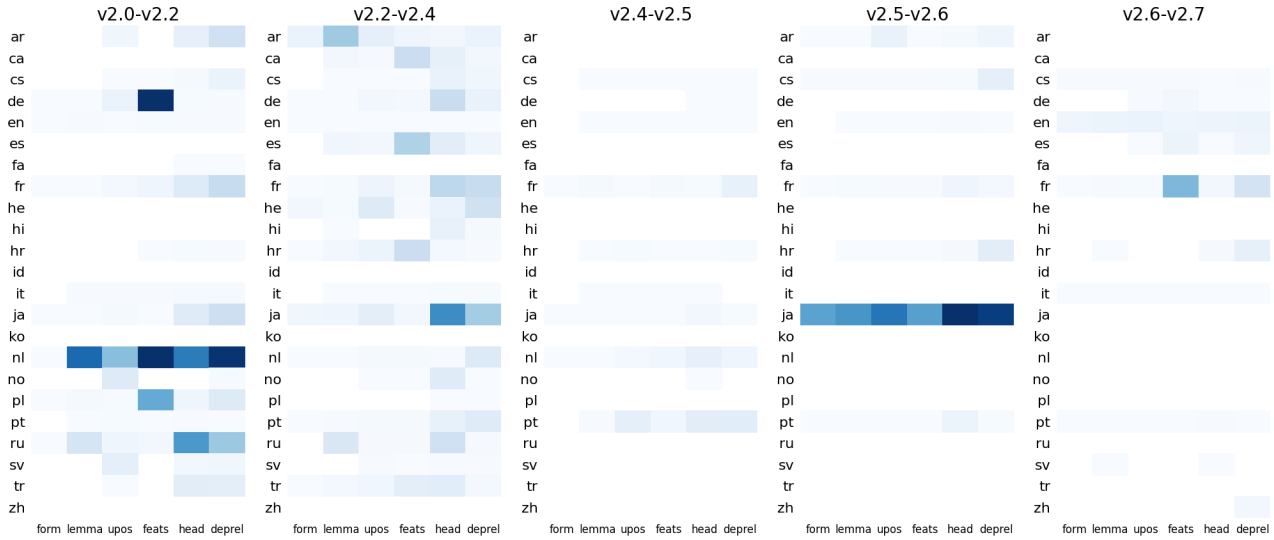


図 1　各言語，バージョンにおける UD コーパスの更新割合．色が濃いほど更新割合が高い．

の単語が VERB に変更され (“continuar” (‘continue‘), “deixar” (‘leave‘))，フランス語 (UD2.1) でも AUX タグが “être” (‘be’), “avoir” (‘have’), “faire” (‘do’, ‘make’), “pouvoir” (‘can’) のみに限定され，他の動詞は VERB とタグ付けされた．このように言語間でのタグ付けの一貫性が向上する変化が観察された．

3　実験

我々は異なるバージョンの UD コーパスで訓練した構文解析器で評価表現抽出 [6] を行い，UD コーパスの変化が応用タスクに及ぼす影響を分析した．

3.1　構文解析器の再学習

UD 準拠の構文解析器として UDPipe 1.2 [5] を用いた．UD2.0–2.5 の構文解析器は UDPipe が配布する訓練済みモデル[2] を用い，配布モデルが存在しない UD2.6, UD2.7 は UD2.5 と同じパラメータを用いて訓練した．簡体字中国語のコーパス UD2.0, UD2.2 は存在しないため評価から外した．構文解析器の性能評価として LAS (Labeled Attachment Score) を用いた．

3.2　構文構造に基づく評価表現抽出

評価表現抽出では構文木をトップダウンにたどり，極性辞書と符合させながら好評不評の表現とその対象を抽出する．依存関係やラベルを見ることによって，否定の表現や対象を抽出するための格を正確に捕捉できる仕組みになっている．詳細は [6] を参照されたい．

2)　http://ufal.mff.cuni.cz/udpipe

3.3　評価データと評価指標

評価データは先行研究 [6] と同様に，ホテルや携帯電話の言及単位または文単位の注釈付きレビューを整形し，好評/不評のラベルが付いた文を各言語合計 500 文程度抽出したものを用いた．

システムの適合率は，出力した極性が正解の極性と同一である割合，再現率は正解の極性と同じ極性を持つ評価表現が検出された割合である．総合的な評価指標として，適合率を重視した F2 値（式 1 で $\beta = 2$ としたもの）を用いる．一般的な F1 値 ($\beta = 1$) を用いると，文法構造を意識せずに好評・不評を含む単語を抽出するアプローチによる数値が高くなるものの，適合率が低く実用性と乖離する．

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

4　結果と考察

本節では多言語の構文解析と評価表現抽出の性能の関係を定量的，定性的に評価する．

4.1　構文解析の性能と評価表現抽出の関係

図 2 に構文解析 (LAS) と評価表現抽出 (F2) の関係を示す[3]．言語ごとにデータセットの作成元が異なるため言語をまたいだ絶対的な数値の比較は難しいが，データセットの分野が原因で評価表現抽出が難しい言語 (韓国語など) を除けば，言語間の F2 と LAS には大まかな正の相関が存在する．

3)　インドネシア語と韓国語の UD2.0 では lemma が出力されず，辞書との符号ができないため評価には含まない

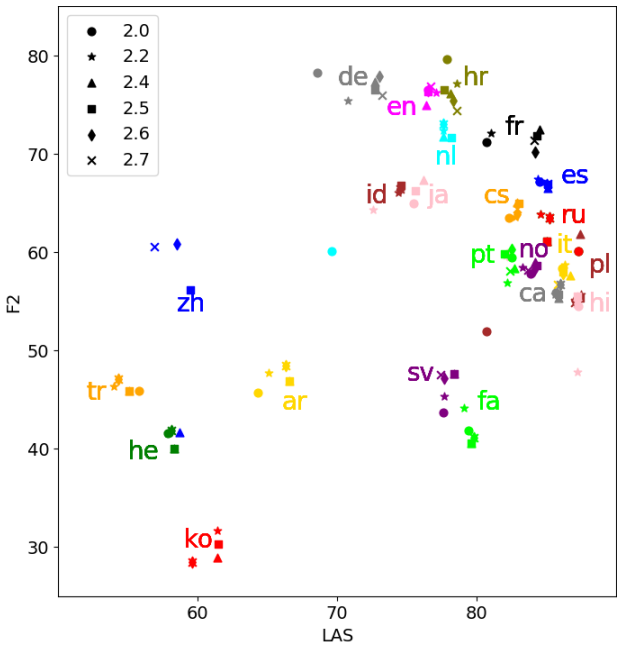
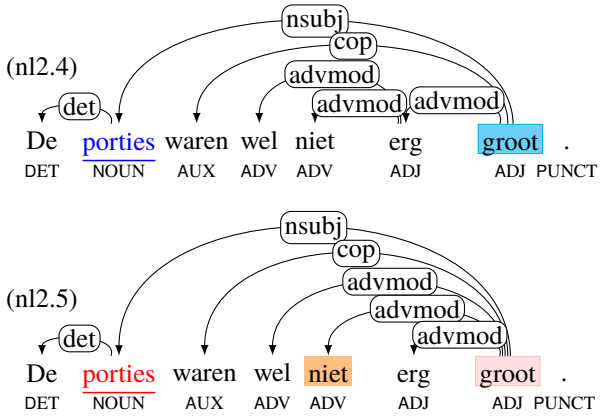


図 2　UD のバージョンごとの LAS と F2 値

一つの言語内でも LAS や F2 値には変化が見られ，その変化は辞書などの評価表現抽出の仕組みではなく，構文解析器の変更によってのみ発生する．F2 値が 40.0 以上で，UD2.0 と UD2.6 の F2 値に 1.0 以上の差がある言語を比較すると，9 言語中 7 言語で UD2.6 のほうが UD2.0 よりも F2 値が高い[4]．つまり UD コーパスの更新に伴い，応用タスクの性能も徐々に向上しているといえる．コーパスの更新に伴う出力の変化については次に述べる．

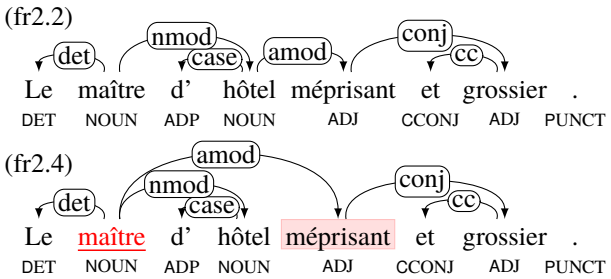
4.2　UD 更新に伴う変化の観察

ここでは UD のバージョンの変化によって評価表現抽出の結果が変わった例について紹介する．まず例として挙げる構文解析と評価表現の表記について説明する．

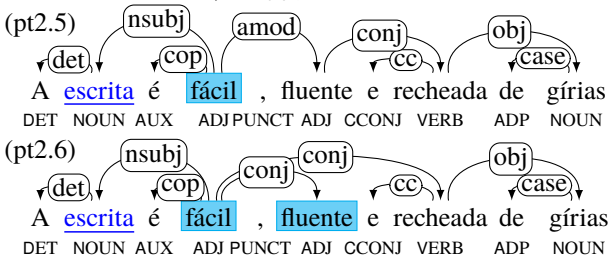


4）日本語 UD2.6 では語の単位が変更されたため F2 値が低い (nl2.4) はオランダ語の UD2.4 で訓練した UDPipe に文を入力し，評価表現を抽出した結果である．ハイライトのうち青が好評表現，赤は不評表現を表し，下線は好評/不評表現の対象を表す．オレンジは極性を反転させる否定表現である．

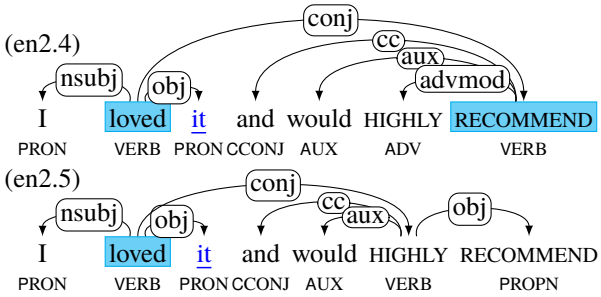
(nl2.4) では，否定の副詞 “niet”(‘not’) が評価表現 “groot”(‘large’) に直接係っておらず文全体として極性誤りが生じているが，(nl2.5) では “groot” と “niet” の依存構造が正しいため，否定による極性の反転を検出できている．



フランス語では UD2.4 で多数の依存関係ラベルや PoS タグが更新され，構文解析の性能が向上している．(fr2.4) の名詞句では，形容詞 “méprisant” (‘contemptuous’) が head の名詞 “maître” (‘master’) にかかっているため，不評表現が正しく検出された．



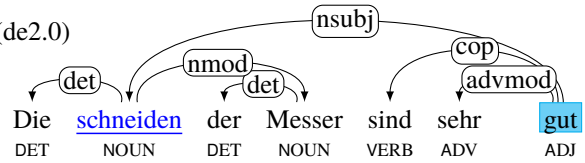
評価表現抽出では並列構造を示す conj ラベルが重要な働きをする．例えば，(pt2.5), (pt2.6) ではいずれも root は “fácil” (‘easy’) だが，(pt2.6) では “fluente” (‘fluent’) に conj のラベルが付いているため，“fluente” も極性表現として正しく抽出されている．



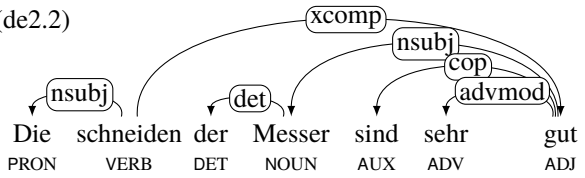
実際のレビューの中には文体が崩れた文も出現する．英語を例にとると，(en2.5) では，通常大文字で書かれない “HIGHLY RECOMMEND” を PROPN タ

グと誤判定する例が存在した．UD で訓練した構文解析器を応用タスクで用いる際には，このような入力に対しても頑健であることが望ましい．

(de2.0)

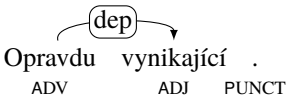


(de2.2)

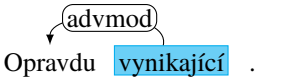


同様の問題はドイツ語でも発生している．ドイツ語では名詞の最初の文字は大文字で記述するが，下の例のように名詞 “Schneiden” (‘sharpness’) が小文字になっていた場合，構文構造上名詞の可能性しか存在しないにもかかわらず，同じ綴りを持つ動詞の不定形と認識される．つまり，英語やドイツ語での品詞判定には大文字であるかどうかが少なからず影響していることがわかる．

(cs2.5)

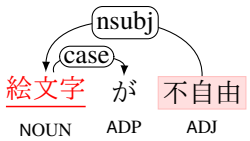


(cs2.6)

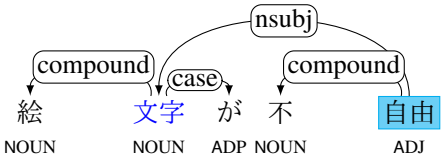


特定できない依存関係を指す dep ラベルを減らすことも UD コーパスの課題の一つである．チェコ語では dep ラベルの比率が他の言語のコーパスよりも大きいため，解析結果に頻繁に出現する．(cs2.6) の出力では正しく advmod のラベルがつけられ，好評表現が抽出できている．

(ja2.5)

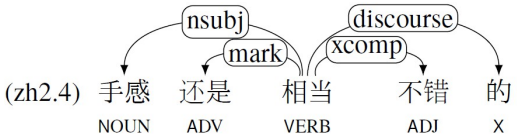


(ja2.6)

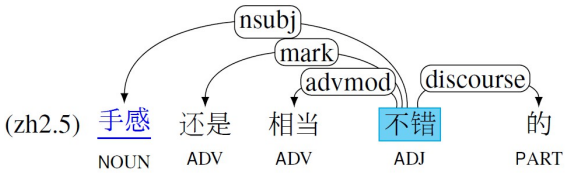


また，単語分割の方針変更は応用タスクに大きな影響を及ぼす．日本語では UD2.6 の単語単位が国語研短単位になったことにより，UD2.5 までは不評表現として 1 つのトークンとして検出されていた “不自由” が，(ja2.6) では “不” + “自由” として検出され，極性が反転している．アノテーション方針の変化に合わせて否定の “不” などの処理をシステムに追加することが正しく極性を検出するために必要である．

(zh2.4)



(zh2.5)



簡体字中国語はリリース時期が今回分析した言語の中では比較的遅く，UD2.5 から正式にリリースされた．そのため，UD2.4 の公開時にプレリリースされ dev ブランチに存在した簡体字コーパスとの比較を行なった．(zh2.4) で存在した “相当” (‘very’) の PoS タグの出力誤りが (zh2.5) で改善され，極性が正しく抽出されるようになった．簡体字中国語は LAS が低い言語のうちの一つであり，今後の改善が期待される．

5 結論

本論文では UD2.0–2.7 での更新を，構文解析器の出力の変化が応用タスクに及ぼす影響という観点から分析した．23 言語の UD コーパスは継続的に更新されており，評価表現抽出の結果はそれぞれのバージョンによって異なることが分かった．言語ごとの出力例から，品詞タグ付けや依存関係ラベルの中でも否定や conj ラベルが評価表現抽出に影響を与えることを示した．スタイルが崩れた文に対する頑健性に関わる問題も見られた．UD の各言語の代表的なコーパス間でドメインが異なり，今回対象としたレビュー文によくみられる，名詞句やスペルミス，大文字小文字や文末のピリオドが無い入力に頑健でない場合が存在した [9]．

UD2.4 以降はいくつかの言語で更新頻度が低下しているが，まだ基本的な構文構造の出力において改善すべき点は多い．UD の継続的な更新への期待と，その他の応用タスクでの UD 準拠の構文解析器の性能評価を今後の展望とする．

参考文献

[1] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, 2020.

[2] Joakim Nivre and Chiao-Ting Fang. Universal dependency evaluation. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies*, pp. 86–95, 2017.

[3] Daniel Zeman, et al. Universal Dependencies 2.6, 2020. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[4] Daniel Zeman, et al. CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–19, 2017.

[5] Milan Straka and Jana Straková. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 88–99, 2017.

[6] Hiroshi Kanayama and Ran Iwamoto. How Universal are Universal Dependencies? Exploiting Syntax for Multilingual Clause-level Sentiment Detection. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 4063–4073, 2020.

[7] 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.

[8] 金山博, 岩本蘭. 多言語評価表現抽出を通じた universal dependencies の検証. 言語処理学会第 26 回年次大会予稿集, 2020.

[9] 金山博, 岩本蘭, 村岡雅康, 大湖卓也, 宮本晃太郎. 名詞句の処理に頑健な構文解析器. 言語処理学会第 27 回年次大会予稿集, 2021.

A 付録

A.1 評価データ

表 3 に評価表現抽出で使用した評価データの情報を示す．言及単位のアノテーションが付いた多言語の評価データが少ないため，言及単位と文単位のデータをそれぞれ整形して用いた．

A.2 構文解析と評価表現抽出

表 4 に言語，バージョンごとの LAS と F2 値を示す．これは図 2 を表にし，さらに適合率 (prec) と再現率 (rec) の情報を加えたデータである．

表 3 評価データの情報．

言語	分野	好評文	不評文	平均文長 (語)
ar	ホテル	250	250	27.0
ca	ホテル	211	149	14.8
cs	レストラン	250	250	16.5
de	カトラリー	297	62	13.8
en	レストラン	250	250	14.7
es	レストラン	250	250	15.4
fa	携帯電話	250	187	25.6
fr	レストラン	250	250	16.3
he	ニュース	250	250	13.2
hi	携帯電話	250	183	18.1
hr	レストラン	250	67	20.0
id	レストラン	250	250	10.7
it	ホテル	250	250	15.0
ja	携帯電話	238	295	22.8
ko	映画	250	247	9.2
nl	レストラン	250	250	14.9
no	その他	250	250	19.2
pl	香水	250	127	9.5
pt	本	250	250	22.6
ru	レストラン	250	250	17.3
sv	その他	250	250	18.9
tr	レストラン	250	250	10.2
zh	携帯電話	252	248	34.7

表 4 UD のバージョンごとの構文解析 (LAS) と評価表現抽出 (F2) の性能評価．

	UD2.0		UD2.2		UD2.4		UD2.5		UD2.6		UD2.7	
language	LAS	F_2 *p*rec *r*ec	LAS	F_2 *p*rec *r*ec	LAS	F_2 *p*rec *r*ec	LAS	F_2 *p*rec *r*ec	LAS	F_2 *p*rec *r*ec	LAS	F_2 *p*rec *r*ec
ar	64.3	45.7 88.2 15.6	65.1	47.7 85.6 17.2	66.6	46.9 83.7 17.0	66.6	46.9 83.7 17.0	66.3	48.4 87.4 17.4	66.3	48.4 87.4 17.4
ca	85.7	55.9 89.0 22.5	85.6	55.7 90.8 21.9	85.9	55.3 87.1 22.5	85.9	55.7 87.2 22.8	86.0	56.7 89.2 23.1	86.0	56.7 89.2 23.1
cs	82.3	63.5 86.1 31.0	82.8	63.6 87.8 30.2	82.9	65.0 87.2 32.2	83.0	64.9 87.4 32.0	82.9	63.8 87.1 30.8	82.9	64.4 87.7 31.2
de	68.6	78.2 93.6 47.1	70.8	75.4 92.7 43.2	72.7	77.3 92.7 46.5	72.7	76.5 90.3 47.4	73.0	77.9 92.0 48.2	73.2	75.9 91.8 44.8
en	76.5	76.5 93.2 44.6	77.1	76.2 92.6 44.6	76.4	74.9 91.3 43.6	76.5	76.3 93.5 44.0	76.5	76.4 92.5 45.0	76.7	76.9 93.3 45.2
es	84.5	67.1 88.7 34.0	84.4	67.4 89.6 33.8	85.1	66.5 90.3 32.4	85.1	66.9 90.4 32.8	85.0	66.9 89.2 33.4	85.1	67.0 89.2 33.6
fa	79.4	41.8 83.1 14.0	79.1	44.1 83.3 15.3	79.6	40.6 81.4 13.5	79.6	40.6 81.4 13.5	79.8	41.1 79.7 14.0	79.8	41.1 79.7 14.0
fr	80.7	71.2 89.5 39.2	81.0	72.1 90.5 39.8	84.5	72.5 90.9 40.0	84.3	71.8 90.3 39.4	84.2	70.2 89.7 37.6	84.1	71.4 89.6 39.4
he	57.9	41.6 82.1 14.0	57.9	41.6 82.1 14.0	58.3	40.0 83.1 13.0	58.3	40.0 83.1 13.0	58.1	41.8 87.0 13.6	58.1	41.8 87.0 13.6
hi	87.3	54.5 83.3 22.9	87.2	47.8 81.7 18.0	87.2	55.5 88.1 22.4	87.2	55.5 88.1 22.4	87.2	54.8 88.5 21.7	87.2	54.8 88.5 21.7
hr	77.9	79.6 95.4 47.9	78.6	77.1 94.0 44.8	78.1	76.1 92.6 44.5	77.7	76.5 94.4 43.5	78.3	75.4 93.0 42.9	78.6	74.4 91.2 42.9
id	74.3	- - -	74.4	66.0 93.2 30.4	74.5	66.7 92.4 31.6	74.6	66.8 92.9 31.4	74.5	66.4 93.3 30.8	74.5	66.4 93.3 30.8
it	86.1	58.3 85.7 25.6	86.3	58.7 89.1 24.8	86.7	57.6 85.4 25.0	86.2	58.0 86.6 25.0	86.2	57.8 85.9 25.0	85.8	56.7 87.2 23.6
ja	75.5	64.9 92.0 29.8	72.6	64.3 90.5 29.8	76.2	67.3 92.0 32.5	75.6	66.2 90.9 31.7	82.2	59.7 88.0 26.1	82.2	59.7 88.0 26.1
ko	60.5	- - -	61.4	31.7 83.3 9.1	61.4	28.9 83.3 8.0	61.5	30.3 84.0 8.5	59.6	28.5 84.8 7.8	59.6	28.5 84.8 7.8
nl	69.6	60.1 90.6 25.6	77.6	72.4 90.7 40.0	77.6	71.7 89.4 40.0	78.2	71.6 90.8 38.8	77.6	73.0 91.2 40.6	77.6	73.0 91.2 40.6
no	83.9	57.8 77.5 28.6	83.3	58.4 77.9 29.2	84.2	59.0 77.4 30.2	84.3	58.6 76.6 30.2	84.1	58.2 77.5 29.2	83.7	58.2 77.5 29.2
pl	80.7	51.9 91.0 19.1	87.5	55.7 95.0 21.0	87.4	61.8 90.9 27.1	87.4	55.3 94.9 20.7	87.0	55.0 93.8 20.7	87.0	54.9 93.8 20.7
pt	82.5	59.4 80.5 29.0	82.2	56.9 79.6 26.6	82.7	58.3 79.1 28.4	82.0	59.8 81.8 28.8	82.5	60.3 83.3 28.6	82.4	58.1 80.7 27.4
ru	87.3	60.1 88.8 26.2	84.6	63.8 88.8 30.0	85.0	61.1 87.3 27.8	85.0	61.1 87.3 27.8	85.2	63.5 89.4 29.4	85.2	63.5 89.4 29.4
sv	77.6	43.7 74.8 16.4	77.7	45.3 78.8 16.8	78.4	47.6 81.0 18.0	78.4	47.6 81.0 18.0	77.7	47.2 81.4 17.6	77.4	47.5 79.5 18.2
tr	55.8	45.9 92.6 15.2	54.0	46.3 94.9 15.2	55.1	45.9 94.9 15.0	55.1	45.9 94.9 15.0	54.3	47.1 95.1 15.6	54.3	47.1 95.1 15.6
zh	-	- - -	-	- - -	58.7	41.7 86.1 13.6	59.5	56.1 84.4 24.0	58.5	60.8 87.5 27.4	56.9	60.5 86.8 27.4