Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 7078–7087
Marseille, 20-25 June 2022

The Persian Dependency Treebank Made Universal

Mohammad Sadegh Rasooli[1]*, Pegah Safari[2]*, Amirsaeid Moloodi[3]*, Alireza Nourian[4]
[1]Microsoft, Mountain View, CA, USA
[2] Faculty of Computer Science and Engineering, Shahid Beheshti University, Tehran, Iran
[3] Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran
[4] Sobhe, Tehran, Iran
[1]mrasooli@microsoft.com, [2]p_safari@sbu.ac.ir,
[3]amirsaeid.moloodi@shirazu.ac.ir, [4]nourian@sobhe.ir

Abstract

We describe an automatic method for converting the Persian Dependency Treebank (Rasooli et al., 2013) to Universal Dependencies. This treebank contains 29107 sentences. Our experiments along with manual linguistic analysis show that our data is more compatible with Universal Dependencies than the Uppsala Persian Universal Dependency Treebank (Seraji et al., 2016), larger in size and more diverse in vocabulary. Our data brings in labeled attachment F-score of 85.2 in supervised parsing. Also, our delexicalized Persian-to-English parser transfer experiments show that a parsing model trained on our data is ≈2% absolutely more accurate than that of Seraji et al. (2016) in terms of labeled attachment score.

Keywords: Universal Dependencies, Persian Dependency Treebank, Automatic Conversion

1. Introduction

In recent years, there has been a great deal of interest in developing universal dependency treebanks (McDonald et al., 2013; Rosa et al., 2014; Nivre et al., 2020). The main goal of the Universal Dependencies project (Nivre et al., 2020) is to develop a consistent linguistic annotation scheme in different levels from tokenization to syntactic dependency relations. As a result, the majority of annotation discrepancies disappear, and the resulting dataset facilitates several cross-lingual natural language processing tasks including part-of-speech transfer (Täckström et al., 2013), syntactic transfer (Naseem et al., 2010; McDonald et al., 2011; Ammar et al., 2016; Zhang et al., 2019), and probing (Tenney et al., 2019; Hewitt and Manning, 2019). Starting with 10 treebanks in 2015, there are 217 treebanks in version 2.9 (November 2021).

Persian (aka Farsi) is a pro-drop morphologically rich language with a high degree of free word order and a unique light verb construction (Karimi-Doostan, 2011). Despite its importance, it still suffers from lack of sufficient annotated data. The Uppsala Universal treebank (Seraji et al., 2016) is currently the only publicly available universal treebank for Persian. It is a valuable resource based on news genre, and has been used as a testbed in previous work (Zeman et al., 2018; Chi et al., 2020). Among other non-universal treebanks, the Persian dependency treebank (PerDT) (Rasooli et al., 2013) is significantly larger than (Seraji et al., 2016) (29K vs. 6K sentences), and its sentences are sampled from contemporary Persian texts in different genres (as opposed to only news genre).

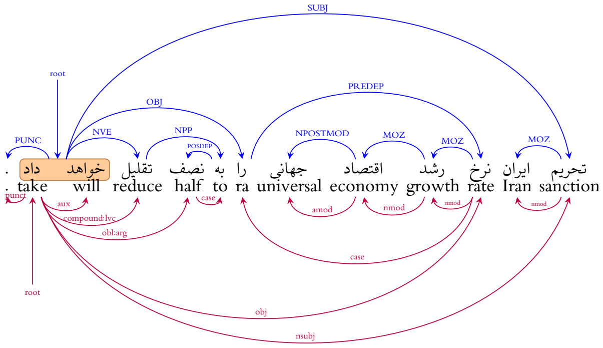

SUBJ
root
OBJ
PREDEP
PUNC
NVE
NPP
POSDEP
NPOSTMOD
MOZ
MOZ
MOZ
. داد خواهد تقلیل نصف به را جهانی اقتصاد رشد نرخ ایران تحریم
. take will reduce half to ra universal economy growth rate Iran sanction
punct
aux
compound:lvc
obl:arg
case
amod
nmod
nmod
nmod
case
root
obj
nsubj

Figure 1: An example of our automatic conversion. The converted universal labels are shown at the bottom of words in red while the original relations are in blue. Translation: *Sanctions on Iran will halve the world's economic growth rate.*

In this paper, we propose an automatic method for converting PerDT (Rasooli et al., 2013) to Universal Dependencies (An example of such conversion is shown in Figure 1). After a thorough analysis of dependency relations in the treebank, we design different mapping rules to generate trees with universal relations. This process involves a series of steps including unifying tokenization, part-of-speech tags, named-entity recognition, and finally mapping dependencies. The mapping for many relations are not necessarily one-to-one, and we have to deal with peculiar cases that are specific to

*Rasooli and Safari equally contributed in the conversion and experimentation process. Rasooli and Moloodi equally contributed in the linguistic design of conversion rules and manual investigation of conversions.

certain structures in modern Persian. Therefore, our approach is neither a blind one-to-one mapping, nor an expensive and time-consuming manual process. We empirically show that our annotations are more compatible with the Universal guidelines via learning a delexicalized transfer model with more than 2% absolute difference in labeled attachment score. The summary of our contributions is as follows:

- We propose an automatic annotation conversion process with manual care of special cases. We develop a new Persian Universal Treebank with 29107 sentences which is nearly 5 times larger than the treebank of Seraji et al. (2016) with 5997 sentences.

- We develop a modified and corrected version of PerDT with the Universal tokenization scheme. Moreover, the new release resolves various tagging errors in the original dataset. Most of these corrections are made by manually fixing annotation errors flagged by our mapping pipeline.

2. Related Work

There has been a great deal of interest in designing and developing Persian dependency treebanks (Pouramini and Mozayani, 2007; Seraji et al., 2012; Seraji et al., 2014; Seraji et al., 2016; Rasooli et al., 2011b; Rasooli et al., 2013; Ghayoomi and Kuhn, 2014). Among them, the Uppsala UD treebank (Seraji et al., 2016) is the only treebank with Universal Dependencies. We have found some caveats in the Uppsala Universal Treebank (Seraji et al., 2016). This causes annotation discrepancies in some frequently used dependency relations such as *compound:lvc*, *cop*, *csubj*, *fixed*, *obl*, and *xcomp* (see §4.3 for more details).

We primarily focus on converting the Persian dependency treebank (PerDT) (Rasooli et al., 2013). PerDT has been used in previous studies for Persian dependency parsing (Khallash et al., 2013; Feely et al., 2014; Nourian et al., 2015; Pakzad and Minaei-Bidgoli, 2016). It has been extended to other representations including semantic roles (Mirzaei and Moloodi, 2016) and discourse (Mirzaei and Safari, 2018). It is also included in the HamleDT collection (Rosa et al., 2014).

3. Approach

We decompose the conversion process into four steps: 1) tokenization, 2) part-of-speech mapping, 3) systematic changes to PerDT, and 4) dependency relation mapping.

3.1. Tokenization

There are two key differences between tokenization in PerDT and UD:

1. Multiword inflections of simple verbs in PerDT are grouped as one word with spaces between parts following the deterministic rules from (Rasooli et al., 2011a). For UD tokenization, we use the guidelines in (Rasooli et al., 2013, Table 3) to find the main verb in each expression and make the other parts as “aux” dependent of the main verb. It results in introducing the “AUX” part-of-speech tag and “aux” dependency relation (“aux:pass” for passive verbs).

2. Clitics in PerDT are only detached from words in cases for which they play an object or verbal role. Other clitics are pronominal clitics attached to nouns, prepositions, pronouns and adjectives. By looking at the word lemma, we recover those pronouns, detach them, and assign their heads to the closest nominal word with the “MOZ” (*Ezafe*) dependency label. An example of such change is “كتابش” [ketabæʃ] (his book) which is tokenized into two words “كتاب” [ketab] (book) and “ش” [ʃ] (his).

3.2. POS Mapping

This is the most straightforward step except for proper nouns. Table 1 shows the mappings. We could only discover a small portion of proper nouns by finding noun phrases with an identifier (IDEN POS tag for words such as “Dr.” or “Mr."). In addition to mapping the IDEN POS to PROPN, we use a recent BERT-based Persian named-entity tagger (Taher et al., 2020) to recover additional proper nouns. The tagger can find 7 different entities including date, location, money, organization, percent, person and time. We only consider the *person* and *location* entities, and manually revise the results to add missing entities, foreign words and the name of months. We have also tried our best to correct wrong detects through our revision.

3.3. Systematic Changes to the Original Annotation

There are a few systematic decisions in PerDT that we believe our suggestions are better fit for it. Before starting to convert the treebank to Universal Dependencies, we have made the following systematic changes to it:

- We convert the order of verbal conjunctions in the original data. In PerDT, verbal conjunctions are conventionally attached from the end to the beginning (Dadegan Research Group, 2012).[1] We find this convention unintuitive and reverted the order of conjunctions. Figure 2 shows an example of such rotation.

- Words such as “billion”, “million”, and “thousand” are tagged as nouns.[2] This might be due to the fact that these words can be inflected as plurals while numbers cannot be inflected in Persian. We believe that a better tagging decision

[1] Examples in `https://bit.ly/2Mfz1iH`

[2] Examples in `https://bit.ly/2Y105Yv`

PerDT	Condition	UD
V		VERB
N	NER=False	NOUN
	NER=True	PROPN
SUBR		SCONJ
CONJ		CCONJ
ADV		ADV
ADJ	NER=False	ADJ
	NER=True	PROPN
PR		PRON
PUNC		PUNCT
ADR		INTJ
IDEN		PROPN
PART	Word=را	ADP
	Word∈{آخر,خوب}	INTJ
	Otherwise	PART
PREM		DET
PRENUM	Cardinal	NUM
	Ordinal	ADJ
PREP POSTP		ADP
POSTNUM	Cardinal	NUM
	Ordinal	ADJ
PSUS		INTJ

Table 1: Mapping rules for part-of-speech tags. For each tag in PerDT (first column) based on the condition in the second column, it is mapped to the corresponding UD tag (third column). Relations without condition are mapped in a straightforward way.

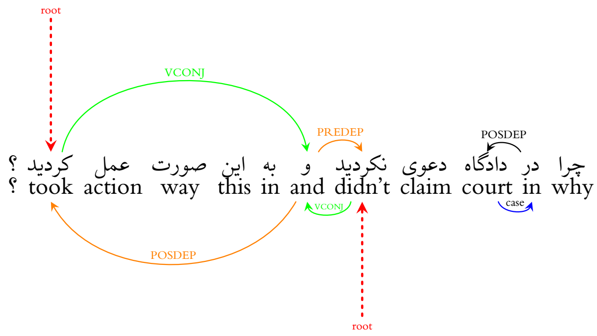

root
VCONJ
PREDEP
POSDEP
چرا در دادگاه دعوی نکردید و به این صورت عمل کردید ؟
? took action way this in and didn't claim court in why
VCONJ
case
POSDEP
root

Figure 2: The result of applying rotations of conjunctions for the sentence *Why didn't you defend in the court **and** acted like that?*. In this example, case rotation for preposition is also shown.

for these words is *number* since their inflection as plurals is due to a special kind of *zero derivation* or *conversion* numbers to nouns in particular contexts (Booij, 2012).

- PerDT assumes that all inflections of "شدن" [ʃodæn] is passive and its lemma is "کردن" [kædæn]. We have changed this assumption and use the superficial lemma for those instances. The decision makes our data similar to the annotations of Seraji et al. (2016).

Correction Type		#	%
Lemma	Systematic	3694	0.762
	Others	59	0.012
POS	Systematic	529	0.109
	Others	298	0.061
FPOS	Systematic	3693	0.762
	Others	90	0.018
Dependency head	Systematic	27407	5.658
	Others	967	0.199
Dependency label	Systematic	18516	3.823
	Others	656	0.135
Word Form		39	0.008

Table 2: Statistics of PerDT corrections. By systematic, we mean deterministic corrections such as verbal conjunctions (see §3.3 for details).

Table 2 shows the statistics of all changes we made including systematic ones or the manual fixes for incorrect annotations.

3.4. Dependency Relation Mapping

PerDT contains 43 syntactic relations. Compared to UD scheme in which content words are considered as heads, PerDT assigns prepositions as the head of prepositional phrases and auxiliary verbs as the root of sentence. Before applying the conversion rules, we label words that are not well-edited and typed as more than one token as *goeswith*. We then label proper noun phrases that are not syntactically compositional as *flat:name*. We also analyze complex numbers as *flat:num* and their coordinating conjunctions as *cc* dependent of each following word. Afterwards we follow the rules in Table 3. As depicted in the Table, there are conditions that should be satisfied before applying a conversion, and some actions such as flipping a head with its dependent are needed before certain mappings. Finally, we label the few remaining undecided dependencies as *dep*.

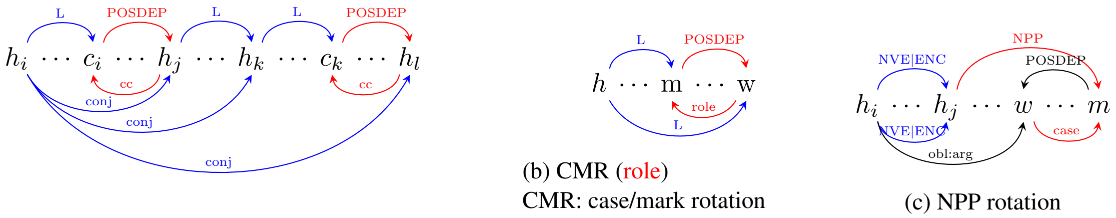
Figure 4 shows three examples of our conversion for which we highlight the most challenging rotations. In the first tree (4a), we show how we create the "flat:num" dependency as well as copula conversion. In the second example (4b), we show how our tokenization of multi-word verbs works. In this example, an example of the "compound:lvc" dependency is also shown. Moreover, we see a rare case of *iobj* in this example for which comes from a second object role in Persian. In the third tree (4c), we see an Ezafe dependent with an attached modifier for which we tokenize it and assign a "nmod" label. In all of these trees, upward relations are dependencies in PerDT while downwards are their equivalent universal relations shown in the same color.

PerDT label	Precondition	Pre-action	UD
ACL, PRD		CMR (mark) [fig.3b]	ccomp
ADV	?→{ADJ,ADV} Otherwise		advmod obl
AJCONJ, AVCONJ, NCONJ, PCONJ, VCONJ		Conj rotation [fig.3a]	conj
AJUCL		CMR (mark) [fig.3b]	advcl

APOSTMOD
APREMOD
	?→{ADP}
	?→{NUM}
	?→*nominal*
	?→{ADJ}
	?→DET
	?→{ADV}
		CMR (case) [fig.3b]	obl
			nummod
			nmod
			amod
			det
			advmod

ADVC, AJPP, NEZ, VPP, VPRT		CMR (case) [fig.3b]	obl:arg
APP			appos
COMPPP	∃ dep Otherwise	CMR (case) [fig.3b]	case fixed
NEZ			
ENC, NE, NPRT, NVE		CMR (case) [fig.3b]	compound:lvc
LVP			compound:lv
NCL	?→SCONJ Otherwise	CMR (mark) [fig.3b]	acl
MESU		Dep →Head (Flip)	nmod
MOS	AUX→? Otherwise	Dep →Head (Flip)	cop xcomp
MOZ	?→NOUN ?→ADJ Otherwise	CMR (case) [fig.3b]	nmod amod advmod
NADV	?→NOUN ?→ADJ Otherwise	CMR (case) [fig.3b]	nmod amod advmod
NPOSTMOD			amod
NPP	{NVE|ENC}→? Otherwise	NPP rotation [fig.3c] CMR (case) [fig.3b]	obl:arg nmod
NPREMOD	?→DET ?→ cardinal Otherwise		det nummod amod
OBJ	∃ OBJ2 sib. Otherwise		iobj obj
OBJ2			iobj
PARCL	?→CCONJ Otherwise	Dep →Head (Flip)	conj parataxis
PART			mark
PUNC			punct
PROG	Active→? Passive→?		aux aux:pass
ROOT			root
SBJ	Active→? Passive→?		nsubj nsubj:pass
TAM			xcomp
VCL	Modal verb→? ∃ MOS,∄ SUBJ sib. Otherwise	Dep →Head (Flip) CMR (mark) [fig.3b]	aux csubj ccomp
PREDEP	NUM→? NOUN→PRON ?→CCONJ ?→NOUN *Last mapping*		advmod dislocated cc obl advmod
POSDEP	NOUN→{هم، نیز} ?→NOUN ?→CCONJ *Last mapping*		dep obl cc advmod

Table 3: Mapping rules for dependencies. PerDT dependency labels are described in Rasooli et al. (2013, Table 2). First *Preconditions* (2nd column) should be satisfied. Afterwards, *Preactions* (3rd column) are applied before applying the UD conversions (4th column). These preactions are depicted in Figure 3.

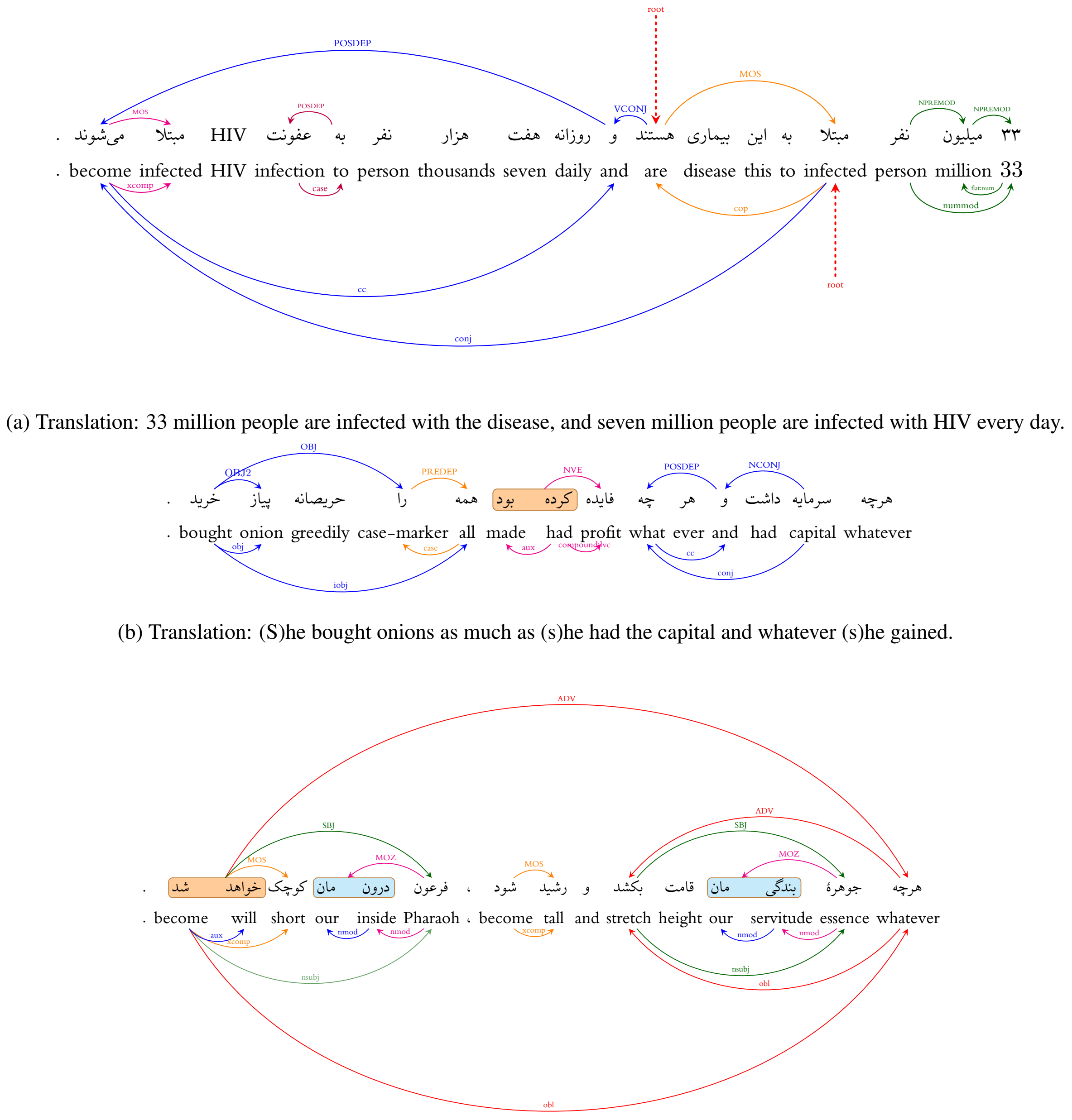


(a) Conj Rotation

(b) CMR (role)
CMR: case/mark rotation

(c) NPP rotation

Figure 3: A graphical depiction of rotation rules used in this work (see Table 3 for their use cases).



POSDEP
root
MOS
MOS
POSDEP
VCONJ
NPREMOD
NPREMOD
۳۳ میلیون نفر مبتلا به این بیماری هستند و روزانه هفت هزار نفر به عفونت HIV مبتلا می‌شوند .
. become infected HIV infection to person thousands seven daily and are disease this to infected person million 33
xcomp
case
flat:num
nummod
cop
cc
root
conj

(a) Translation: 33 million people are infected with the disease, and seven million people are infected with HIV every day.

OBJ
OBJ2
PREDEP
NVE
POSDEP
NCONJ
هرچه سرمایه داشت و هر چه فایده کرده بود همه را حریصانه پیاز خرید .
. bought onion greedily case-marker all made had profit what ever and had capital whatever
obj
case
aux
compound:lvc
cc
iobj
conj

(b) Translation: (S)he bought onions as much as (s)he had the capital and whatever (s)he gained.

ADV
SBJ
MOS
MOZ
MOS
ADV
SBJ
MOZ
هرچه جوهرهٔ بندگی مان قامت بکشد و رشید شود ، فرعون درون مان کوچک خواهد شد .
. become will short our inside Pharaoh ، become tall and stretch height our servitude essence whatever
aux
xcomp
nmod
nmod
xcomp
nmod
nmod
nsubj
nsubj
obl
obl

(c) Translation: As the essence of our lives grows and matures, the Pharaoh within us will shrink.

Figure 4: Examples of our conversions for which they require a lot of structure change to become aligned with the Universal Dependencies annotation guidelines.

Part	Treebank	Sentences#	Tokens#	Types#		
				Word	Lemma	Verb
Train	UDT	4798	122K	13.9K	6.7K	1226
	Ours	**26196**	**459K**	**34.9K**	**20.7K**	**5275**
Dev	UDT	599	15K	3.9K	2.0K	278
	Ours	**1456**	**26K**	**7.0K**	**5.2K**	**1427**
Test	UDT	600	16K	3.9K	3.1K	385
	Ours	**1455**	**24K**	**6.7K**	**5.1K**	**1671**
All	UDT	5997	154K	15.8K	7.6K	1387
	Ours	**29107**	**509K**	**36.7K**	**21.6K**	**5413**

Table 4: Statistics of our data vs. UDT (Seraji et al., 2016) in different data splits.

Test Data	PerDT (Ours)			Seraji et al. (2016)		
ID tagger	×	✓		×	✓	
ID parser	×	×	✓	×	×	✓
Tokens	99.9	**99.99**		100	100.0	
Words	99.1	**99.64**		**99.7**	99.59	
UPOS	82.9	**96.11**		81.9	**95.75**	
Lemmas	80.7	**96.20**		90.2	**89.55**	
UAS	71.2	71.2	**88.4**	69.5	69.8	**83.5**
LAS	64.4	62.6	**85.2**	62.1	61.0	**79.4**
CLAS	59.9	59.3	**81.6**	56.9	56.1	**74.8**
MLAS	49.5	54.5	**78.9**	46.0	53.9	**73.0**
BLEX	44.6	56.9	**78.2**	52.1	49.0	**65.5**

Table 5: Parsing results based on the CoNLL shared task 2018 (Zeman et al., 2018) evaluation. ID stands for in-domain for which the same training set is used for training a UDPipe model (Straka and Straková, 2017)

4. Experiments and Analysis

The general statistics of our data vs. the Uppsala treebank (Seraji et al., 2016) are shown in Table 4. We observe that our data is superior in many aspects including size and diversity compared to the Uppsala Treebank (Seraji et al., 2016). The most important fact about PerDT is that its sentences are intentionally sampled with the purpose of covering almost all verbs from the Verb Valency Lexicon (Rasooli et al., 2011b) leading to 3.9 times more verb lemmas than the Uppsala Treebank. Table 7 shows the counts of each dependency label in the converted data.

4.1. Supervised Parsing

We evaluate the resulting data by training UDPipe V.2 (Straka and Straková, 2017) along with the pretrained fastText (Grave et al., 2018) embeddings on our data. We also evaluate our models on the Uppsala treebank (Seraji et al., 2016). Table 5 shows the parsing results using a trained model on our data and the Uppsala Treebank evaluated by the CoNLL 2018 shared task evaluation scripts (Zeman et al., 2018). It is worth noting that the goal of this evaluation is not to show which dataset brings in better parsing accuracy: it is clear that the bigger the dataset is, the higher the accuracy can be. Our goal is to show that there is a significant performance difference between the models trained on the two datasets by using the exact same training pipeline. As shown in Table 5, we see that there is a huge tagging and parsing performance difference when we move across the datasets. There are two possible reasons: domain mismatch, and annotation discrepancy. Our analysis show that annotation discrepancy plays an important role here. As described in §4.3, there are some core incompatibilities between the Uppsala treebank (Seraji et al., 2016) and Universal Dependencies guidelines. Our detailed analysis shows that most of cross-dataset errors come from errors in *nmod*, *obl*, *fixed*, and *xcomp*. This is in fact consistent with our manual analysis in §4.3.

Training Data	UAS	LAS
Uppsalla Universal Treebank	45.37	36.42
PerDT (this paper)	**47.31**	**38.59**

Table 6: Delexicalized parser transfer results for which training is done on the delexicalized Persian treebank, and evaluation is applied on the English test data.

4.2. Delexicalized Model Transfer

One way to verify our claim about increased consistency of our UD conversion with the UD guidelines is to learn a transfer model. *Direct transfer* models learn a parsing model trained on a different language and test it on an unseen language. In this setting, we follow the delexicalized parser transfer approach which have been extensively used in previous work (Zeman and Resnik, 2008; McDonald et al., 2011; Täckström et al., 2012). In delexicalized transfer, given the lack of lexical overlap among different languages, we ignore the lexical features from treebank data by removing lexical entries in the treebank, aka delexicalization. The main assumption here is that the part-of-speech features are strong indicators of how syntactic structures exist for a sentence. In this setting, we train on the delexicalized version of the Persian treebanks, and evaluate the trained model on the English treebank. Therefore, a higher parsing accuracy is a strong indicator of annotation transferablity.

We sample the same number of tokens as of Seraji et al. (2016) from PerDT. Afterwards, we delexicalize both of the treebanks, and learn a parser using the Yara Parser (Rasooli and Tetreault, 2015). We train two models with 15 epochs and evaluate them on the delexicaled test set of the Universal English Web Treebank (Silveira et al., 2014). As shown in Table 6, the model trained on PerDT significantly outperforms the other model by 2% both in unlabeled and labeled attachment score (47.31 vs 45.37 UAS, 38.59 vs. 36.45 LAS). This is a strong indicator that our data is more compatible with the UD annotations.

4.3. Problems in the Universal Annotations of the Uppsala Universal Treebank

We briefly mention some of the problems in the Uppsala Universal Treebank (Seraji et al., 2016):

- Seraji et al. (2016) do not determine the *csubj* label in their analysis. For example, in "lɑzem ʔæst ʔu **beresæd**" (it is necessary for him **to arrive**), it is obvious that what comes after "ʔæst" is the clausal subject of the adjectival sentence predicate "lɑzem". A simple syntactic test supports this viewpoint: one can convert the clausal complement "ʔu **beresæd**" to a noun phrase "residæn-e u" (his arrival). The new phrase plays the *nsubj* role of the sentence. Therefore, the clausal complement of the sentence should be *csubj*. Our converted data contains 682 cases of *csubj*.
- Seraji et al. (2016) considers prepositional and possessive complements of adjectival heads as *nmod* and *nmod:poss* respectively. Their analysis clearly stands in contradiction to UD annotation guideline in which *nmod* is used just for dependents of nominal heads. *obl* is much better suited for these cases.
- Seraji et al. (2016) consider "شدن" [ʃodæn] (to become) as copula. What UD asserts under the *cop* (copula) label is that *"the equivalents of to become are not copulas despite the fact that traditional grammar may label them as such."* Instead, it should be deemed as a verbal predicate and its second complement as *xcomp*.

Label	Frequency	%
case	71118	14.1
conj	23739	4.7
acl	10034	1.9
obl	30737	6.1
punct	44336	8.8
cop	6366	1.2
det	10273	2
advmod	9158	1.8
aux:pass	822	0.1
nmod	59442	11.6
appos	1059	0.2
aux	12886	0.16
amod	22576	4.4
compound:lvc	32339	6.4
nsubj:pass	822	0.1
nsubj	27181	5.4
name:flat	7899	1.5
dep	2035	0.4
cc	21300	4.2
root	29107	5.8
advcl	4228	0.8
obj	19999	3.9
xcomp	4920	0.9
parataxis	82	0.01
ccomp	6945	1.3
obl:arg	21510	4.2
flat:num	607	0.1
nummod	5459	1
mark	11982	2.3
fixed	144	0.02
compound:lv	439	0.08
csubj	682	0.1
vocative	174	0.03
compound	42	0.008
iobj	6	0.001
dislocated	1	0.0001

Table 7: Frequency of each universal label the converted dataset.

- "پیدا کردن" ("peydɑ kærdæn") and " حاصل کردن" ("hɑsel kærdæn") are considered as two-word light verbs (Moloodi and Kouhestani, 2017). We consider the non-verbal part as the first part of the two-word light verb, and use the *compound:lv* label for it (439 cases in PerDT). However, Seraji et al. (2016) annotate the nonverbal elements of these complex predicates as *obj* and considers "peydɑ" as a nonverbal element.
- *iobj* label is absent in (Seraji et al., 2016), most likely due to the low frequency of this syntactic relation. Our converted treebank contains 6 cases of *iobj*. Although it is very rare, we believe these peculiar structures are important for further linguistic studies.
- Proper nouns are not labeled in (Seraji et al.,

2016). Ours covers proper nouns with more than $23K$ tokens.

5. Conclusion

We have introduced our automatic approach in making PerDT (Rasooli et al., 2013) universal. During this process, we have faced different challenges such as annotation errors in the original data, tokenization inconsistencies, lack of named entities, and complications in part-of-speech and dependency label mapping. Due to automatic conversions and potential annotation errors in the original treebank, there is always a chance of some annotation incompatibilities between our treebank and the Universal guidelines. Therefore, we cannot claim that our conversion is perfect. However, our experiments have shown that our data is more compatible with the Universal Dependencies guidelines than Uppsala treebank, the only available universal treebank in Persian (Seraji et al., 2016).

Acknowledgments

We would like to thank Morteza Rezaei-Sharifabadi for his efforts in getting copyright permissions and making our treebank publicly available. We also thank Daniel Zeman for helping us in the process of releasing this treebank in the Universal Dependencies website. We thank the anonymous reviewers for their helpful and constructive feedback.

6. Bibliographical References

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., and Smith, N. A. (2016). Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Booij, G. (2012). *The grammar of words: An introduction to linguistic morphology*. Oxford University Press.

Chi, E. A., Hewitt, J., and Manning, C. D. (2020). Finding universal grammatical relations in multilingual bert. *arXiv preprint arXiv:2005.04511*.

Dadegan Research Group. (2012). Persian dependency treebank, annotation manual and user guide. *Supreme Council of Information and Communication Technology (SCICT), Tehran, Iran*.

Feely, W., Manshadi, M., Frederking, R., and Levin, L. (2014). The CMU METAL Farsi NLP approach. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4052–4055, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Ghayoomi, M. and Kuhn, J. (2014). Converting an HPSG-based treebank into its parallel dependency-based treebank. In *LREC*, pages 802–809.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Karimi-Doostan, G. (2011). Separability of light verb constructions in persian. *Studia Linguistica*, 65(1):70–95.

Khallash, M., Hadian, A., and Minaei-Bidgoli, B. (2013). An empirical study on the effect of morphological and lexical features in Persian dependency parsing. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages*, pages 97–107, Seattle, Washington, USA, October. Association for Computational Linguistics.

McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Täckström, O., Bedini, C., Bertomeu Castelló, N., and Lee, J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.

Mirzaei, A. and Moloodi, A. (2016). Persian proposition bank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3828–3835, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Mirzaei, A. and Safari, P. (2018). Persian discourse treebank and coreference corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, Miyazaki, Japan, May. European Languages Resources Association (ELRA).

Moloodi, A. and Kouhestani, M. (2017). The role of metaphor and metonymy in the semantics of persian adjectival preverbs: A cognitive linguistics approach. *Language Art*, 2(2):91–105.

Naseem, T., Chen, H., Barzilay, R., and Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, Cambridge, MA, October. Association for Computational Linguistics.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. (2020). Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Nourian, A., Rasooli, M. S., Imany, M., and Faili, H. (2015). On the importance of ezafe construction in Persian parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 877–882, Beijing, China, July. Association for Computational Linguistics.

Pakzad, A. and Minaei-Bidgoli, B. (2016). An improved joint model: Pos tagging and dependency parsing. *Journal of AI and Data Mining*, 4(1):1–8.

Pouramini, A. and Mozayani, N. (2007). An annotation scheme for a persian treebank. *Proceedings of Computational Linguistics In the Netherlands, CLIN*.

Rasooli, M. S. and Tetreault, J. (2015). Yara parser: A fast and accurate dependency parser. *arXiv preprint arXiv:1503.06733*.

Rasooli, M. S., Faili, H., and Minaei-Bidgoli, B. (2011a). Unsupervised identification of Persian compound verbs. In *Mexican International Conference on Artificial Intelligence*, pages 394–406. Springer.

Rasooli, M. S., Moloodi, A., Kouhestani, M., and Minaei-Bidgoli, B. (2011b). A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank. In *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 227–231.

Rasooli, M. S., Kouhestani, M., and Moloodi, A. (2013). Development of a Persian syntactic dependency treebank. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 306–314, Atlanta, Georgia, June. Association for Computational Linguistics.

Rosa, R., Mašek, J., Mareček, D., Popel, M., Zeman, D., and Žabokrtský, Z. (2014). HamleDT 2.0: Thirty dependency treebanks stanfordized. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2334–2341, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

Seraji, M., Megyesi, B., and Nivre, J. (2012). Bootstrapping a Persian dependency treebank. *Linguistic Issues in Language Technology*, 7(18).

Seraji, M., Jahani, C., Megyesi, B., and Nivre, J. (2014). A Persian treebank with Stanford typed dependencies. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 796–801, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

Seraji, M., Ginter, F., and Nivre, J. (2016). Universal dependencies for Persian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2361–2365, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., and Manning, C. (2014). A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 2897–2904, Reykjavik, Iceland, May. European Languages Resources Association (ELRA).

Straka, M. and Straková, J. (2017). 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*, pages 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Täckström, O., McDonald, R., and Uszkoreit, J. (2012). Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada, June. Association for Computational Linguistics.

Täckström, O., Das, D., Petrov, S., McDonald, R., and Nivre, J. (2013). Token and type constraints for cross-lingual part-of-speech tagging. *Transactions of the Association for Computational Linguistics*, 1:1–12.

Taher, E., Hoseini, S. A., and Shamsfard, M. (2020). Beheshti-NER: Persian named entity recognition using BERT. *arXiv preprint arXiv:2003.08875*.

Tenney, I., Das, D., and Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy, July. Association for Computational Linguistics.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., and Petrov, S. (2018). 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*, pages 1–21, Brussels, Belgium, October. Association for Computational Linguistics.

Zhang, M., Zhang, Y., and Fu, G. (2019). Cross-lingual dependency parsing using code-mixed TreeBank. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natu-

ral Language Processing (EMNLP-IJCNLP), pages 997–1006, Hong Kong, China, November. Association for Computational Linguistics.