Informal Persian Universal Dependency Treebank

Roya Kabiri, Simin Karimi, Mihai Surdeanu
University of Arizona

Introduction

- Persian has two distinct variants: the literary formal form that has been traditionally used in writing, and the conversational informal form typically used in oral communications.
- The phonological, morphological, and syntactic distinctions between the two variants reveal that these varieties have fundamental differences.
- With the advent of social media, speakers tend to write in the spoken informal variety.
- Current computational models of this language focus mainly on the formal variant.
- The distinctions between the two varieties raise domain adaptation problem: parsers, primarily trained on formal data, may face more unknown tokens and structures when evaluated on informal data, and fail to generalize the basic patterns.

Objective

- This study provides empirical evidence to show that the unique characteristics of informal Persian and the lack of an informal Persian annotated corpus necessitate the development of a dedicated treebank for this variant. **iPerUDT: Informal Persian Universal Dependency Treebank**
- This open-resource treebank is annotated in CoNLL-U format within the Universal Dependencies (UD) scheme.
- This may provide a stepping-stone to reveal the significance of informal variants of languages, which have been widely overlooked in natural language processing tools across languages.

Related Work

Two existing formal Persian UD treebanks:

- Uppsala Persian Universal Dependency Treebank (Uppsala UDT) (Seraji et al., 2016)
- Persian Universal Dependency Treebank (PerUDT) (Rasooli et al., 2020)

Two dependency parsing methods that achieved state-of-the-art performance for English:

- Stanza, a graph-based neural dependency parser (Qi et al., 2020)
- PaT, a sequence model, treating dependency parsing as a sequence tagging problem (Vacareanu et al., 2020)

Data Collection & Annotation

- Informal texts were crawled from open access Persian blogs.
- Raw sentences were fed to the Stanza parser, trained on the formal Uppsala UDT.
- Language-specific guidelines of the Uppsala UDT were followed with modifications (when necessary)
- Errors in tokenization, multiword token expansion, lemmatization, part-of-speech and morphological feature tagging, and dependency parsing were manually corrected.

New Dependency Relations

With the exception of the **proper noun (PROPN)** tag which is missing in the Uppsala UDT, the parts of speech tag set in this treebank is identical to that of Uppsala treebank.

The morphological feature **Typo** was added, as informal texts, unlike formal texts, are prone to typos, particularly because a phoneme may be represented by different letters in this language.

The following language-specific syntactic relations (not included in the Uppsala UDT scheme, presumably because they are among the unique properties of informal Persian) were added.

- **compound:redup**

1) ketab metab
   book RED
   'book and other such things'

- **compound:svc**

2) gereft-æm xabid-æm
   take.PST-1SG sleep.PST-1SG
   'I slept.'

- **discourse**

3) axe ta key mi-xab-i
   INTJ till when ASP-sleep-2SG
   'Until when do you sleep?!!!'

- **discourse:top/foc**

4) una-m pæræstar-æn
   they-ADD nurse-COP.PRS.1SG
   'They are nurses, too.'

- **orphan**

5) Jina ketab xærid væli Nikan dæftær
   Jina book buy.PST.3SG but Nikan notebook
   'Jina bought a book but Nikan a notebook.'

Treebank Statistics

The data is split into training (~34k tokens), development (~10k tokens) and test (~10k tokens) sets.

Number of Sentences	3000
Number of Tokens	54,904
Average Sentence Length	18.3
Number of Unique tokens	10889
Number of Multiword Expressions	4091

Table 1. Basic statistics of iPerUDT.

- Multiword expressions account for 7.44% of all tokens in iPerUDT, whereas they account for only 0.84% and 1.41% of tokens in the Uppsala UDT and PerUDT, respectively.

Inter-Annotator Agreement was computed, using Cohen's Kappa coefficient.

Category	Level of agreement	$\mathcal{K}$ value
Lemma	Almost perfect	0.9798
UPOS	"	0.9808
XPOS	"	0.9851
Head	"	0.9859
Dependency relation	"	0.9717

Table 2. Inter-annotator agreement

References

- Seraji, M., Ginter F. and Nivre, J. (2016). Universal dependencies for Persian. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC), Portorož, Slovenia.
- Rasooli, M. S., Safari P., Moloodi A., and Nourian A. (2020). The Persian dependency treebank made universal. arXive:2009.10205, Sep. 2020.
- Vacareanu, R., Barbosa G. C., Valenzuela-Escarcega M. A., and Surdeanu M. (2020). Parsing as tagging. In Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), 5225-5231.
- Qi P., Zhang Y., Zhang Y., Bolton J., and Manning C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 101-108.

Acknowledgment

We would like to thank Roshan Cultural Heritage Institute and Dr. Malakeh Taleghani Graduate Fellowship in Iranian Studies for their support and financial contribution. We are also grateful to Farzaneh Bakhtiary for her help with the annotation.

Experiments & Results

Stanza and PaT parsers were trained on formal Persian treebanks, Uppsala UDT and PerUDT, then were evaluated on the dev sets of the formal treebanks they were trained on (in-domain) as well as the dev set of iPerUDT (out-domain).

		Uppsala UDT		PerUDT	
Model		UAS	LAS	UAS	LAS
Stanza	ID	91.87	89.65	94.16	92.68
	OD	82.39	75.74	81.59	75.76
PaT	ID	86.32	83.14	91.3	89.14
	OD	80.02	71.46	79.24	72.61

Table 3. Accuracy comparison between Stanza and PaT. ID (In-Domain), OD (Out-Domain)

- Regardless of the treebank it is trained on, Stanza outperforms PaT when evaluated in-domain.
- Using a quadratic algorithm, Stanza generates probability scores for all the possible combinations of dependents and heads in a sentence while PaT is linear and predicts the position of the head for a given token directly from the token, without taking into account the head information. Therefore, it requires a huge amount of data to converge.
- The performance of both parsers substantially decreases when evaluated on informal data.
- The erroneous tokens were mostly associated with the distinctive features of informal Persian, including forms of differential object marking (-ro, -o), free forms of pronominals (mæn, ma) functioning as nominal subject or object, pronominal clitics (-esh, -æm, -eshun, -emun), free and clitic forms of copulas (-e, æst, hæst, bud), interjections (kash) and discourse elements (hæm, ke).
- The dependency relations corresponding to informal properties were most affected by the domain shift. A performance reduction of at least 10.00 was observed in 18 dependency relations.
- For instance, the performance of core arguments, including the nominal subject (nsubj) and direct object (obj), decreases by a maximum of 19.58 and a minimum of 10.9. This can be explained by the fact that although formal Persian demonstrates a strict SOV order, informal Persian exhibits a fair amount of flexibility in word order.