An Empirical Study on the Effect of Morphological and Lexical Features in Persian Dependency Parsing

Mojtaba Khallash, Ali Hadian and Behrouz Minaei-Bidgoli

Iran University of Science and Technology

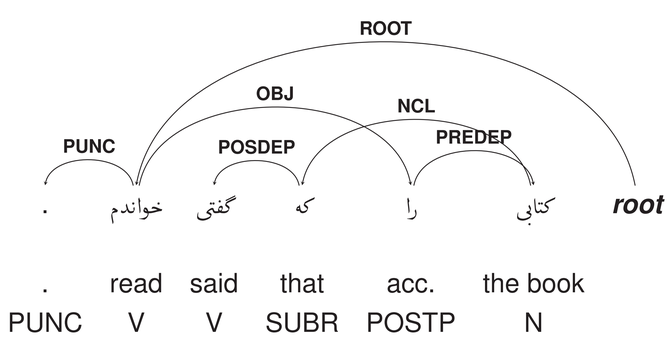
Introduction

- Difficulty of dependency parsing in each language depends on:
  - **free word order**
  - **morphological information**
- Persian is an Indo-European language that is written in Perso-Arabic script (written from right to left).
  - The canonical word order is SOV, but there are a lot of frequent exceptions in word order
  - This language has a high degree of free word order and complex inflections

Corpus

Persian dependency treebank (version 1.0, freely-available at `http://www.dadegan.ir/en`)

- 29,982 sentences
- 498,081 tokens
- 17 coarse-grained POS tags
- 30 fine-grained POS tags
- 22 morphological feature values
- 43 dependency labels
- 21.93% of the sentences are non-projective
- 2.47% of the edges are non-projective



Experiments

In order to study the effects of morphology in dependency parsing of Persian, we organize experiments into three types of challenges and use training and development sets of the treebank.

1. Architecture and Setup

Testing the effects of using automatically derived features, compared to gold features. We use Persian Language Processing (PLP) toolkit and apply it on our training, development, and test sets.

POS tags type	Malt	MST
Gold	87.70	88.04
Predicted	86.98 (-0.72)	86.81 (-1.23)

2. Representation and Modeling

In order to find the best features for the parser, we use ten features:

- **morphological:** Attachment [A], Person [P], Number [N], TMA, Tense [T], and Mood [M]
- **semantic:** Word Clusters [WC], Semantic Verb Clustering [VC], Synset Identifier [SID], and Semantic File [SF]

Feature	Malt	Feature	MST
Baseline	87.70	Baseline	88.04
M	**87.77**	**TMA**	**88.21**$^{+}$
TMA	87.77	M	88.17
T	87.73	P	88.09
SF	87.70	T	88.04
WC	87.69	N	88.04
VC	87.68	SID	88.03
SID	87.67	SF	88.03
A	87.67	WC	88.02
P	87.66	VC	87.98
N	87.65	A	87.93
{**M,SF**}	**87.81**	{**TMA,WC**}	**88.25**

3. Estimation and Smoothing

Suitable way to alleviate the data sparsity problem.

Smoothing	Malt	MST
Baseline	87.70	88.04
Replacing word forms by lemma	87.38	**88.10**
Number Normalization	**87.71**	88.09
Word Clustering	86.98	87.47
Semantic File	87.31	85.25

Final Results

We use the best configurations from the previous section on the training and test sets.

Parser	Method	LAS		UAS		LA	
Malt	Baseline	87.68	(87.04)	90.41	(89.92)	90.03	(89.49)
	Final	**87.91**$^{++}$	(87.16)$^{+}$	90.58^{+}	(90.05)$^{++}$	90.22^{+}	(89.60)$^{+}$
	Diff.	+0.23	(+0.12)	+0.17	(+0.13)	+0.19	(+0.11)
MST	Baseline	87.98	(86.82)	91.30	(90.27)	90.53	(89.90)
	Final	**88.37**$^{++}$	(86.97)	91.55^{++}	(90.36)	90.86^{++}	(90.05)
	Diff.	+0.39	(+0.15)	+0.25	(+0.09)	+0.33	(+0.15)

Baseline and final results of gold (predicted) test set

{khallash,hadian}@comp.iust.ac.ir, b_minaei@iust.ac.ir