PaperDaily -- 2017.09.24

Universal Dependencies are hard to parse – or are they?

Ines Rehbein[♣], Julius Steen[⋆], Bich-Ngoc Do[⋆], Anette Frank[⋆]
Leibniz ScienceCampus
Institut für Deutsche Sprache Mannheim[♣]
Universität Heidelberg[⋆]
Germany
{rehbein,steen,do,frank}@cl.uni-heidelberg.de

Abstract

In the paper, we ask what exactly causes the decrease in parsing accuracy when training a parser on UD-style annotations and whether the effect is similarly strong for all languages.

We show that the encoding in the UD scheme, in particular the decision to encode content words as heads, causes an increase in dependency length for nearly all treebanks and an increase in arc direction entropy for many languages, and evaluate the effect this has on parsing accuracy.

Introduction

Several studies presented experiments on converted trees, offering evidence that a function-head encoding might increase the learnability of the annotation scheme.
Evaluating the learnability of annotation frameworks, however, is not straightforward.
We test the claim that content-head dependencies are harder to parse, using three parsers that implement different parsing paradigms.
We present a conversion algorithm that transforms the content-head encoding of the UD treebanks for coordination, copula constructions and for prepositions into a function-head encoding.

Related work

- Popel et al. (2013) crosslingual investigation of different ways to encode coordination.
- Versley and Kirilin (2015) look at the influence of languages and annotation schemes in UD.
- Gulordava and Merlo (2016) look at word order variation and its impact on dependency parsing of 12 languages.
- Kohita et al. (2017) providing a conversion algorithm for the three functional labels *case*, *dep*, *mark* from the UD scheme.

Conversion algorithm

The phenomena consider in experiments concern the encoding of copula verbs, coordina-tions and adpositions.

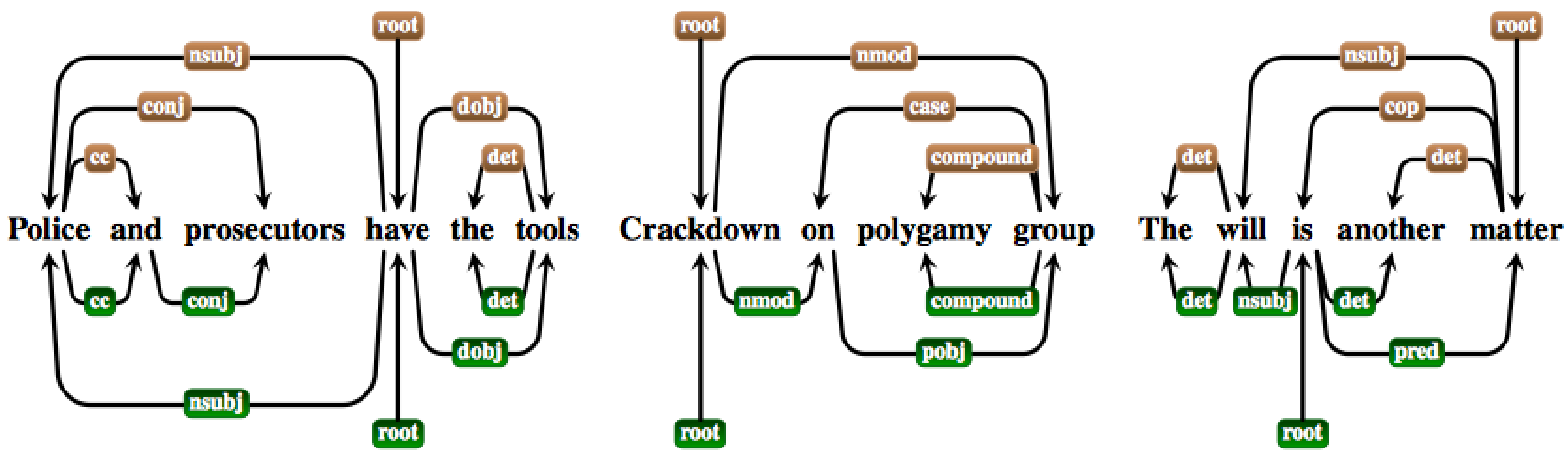

Figure 1: Dependency trees for conversion of coordination (left), prepositions (middle) and copula (right); UD encoding (brown, above) and modified trees with function words as heads (green, below).

Conversion algorithm

			LAS				UAS	% affected
		size	cop	prep	coord	c-p-c	c-p-c	c-p-c
Chinese	**zh**	3,997	100.0	100.0	99.9	99.9	100.0	20.9
Estonian	**et**	14,510	99.9	100.0	100.0	99.9	100.0	23.6
Turkish	**tr**	3,948	99.9	99.8	99.8	99.4	99.8	27.9
Russian-SynTagRus	**ru**	48,171	100.0	100.0	100.0	100.0	100.0	30.6
German	**de**	14,118	99.8	100.0	99.8	99.6	100.0	33.2
Czech	**cs**	68,495	100.0	100.0	99.7	99.7	100.0	35.3
Romanian	**ro**	7,141	99.9	99.9	99.8	99.7	100.0	36.4
English	**en**	12,543	100.0	99.8	99.9	99.6	99.9	37.6
Croatian	**hr**	5,792	100.0	100.0	99.8	99.8	100.0	38.5
French	**fr**	14,554	100.0	99.8	99.9	99.8	99.9	38.5
Catalan	**ca**	13,123	99.9	99.5	99.9	99.4	99.8	38.8
Italian	**it**	12,837	100.0	100.0	99.9	100.0	100.0	40.3
Spanish	**es**	14,187	99.8	99.9	99.9	99.6	99.9	40.3
Bulgarian	**bg**	8,907	100.0	100.0	99.9	99.9	100.0	43.7
Farsi	**fa**	4,798	99.6	100.0	98.8	98.4	100.0	45.7
avg.		*16,475*	*99.9*	*99.9*	*99.8*	*99.6*	*99.9*	*35.4*

Table 1: LAS (excluding punctuation) on the test sets after round-trip conversion for individual transformations and for the combination of all (c-p-c: copula, prep, coord), evaluated against the original UD trees, and UAS for all conversions (c-p-c) (languages are ordered according to the amount of tokens affected by the combination of all conversions; zh: 20.9% – fa: 45.7%).

Experiments

Our main goal is to use the conversion on gold trees in order to compare the impact it has for different languages and thus learn more about how to encode languages with different typological properties to improve monolingual dependency parsing results.

Data `the UD treebanks v1.3`

Three different non-projective parsers

- the graph-based RBG parser (Lei et al., 2014)
- the transition-based IMSTrans parser (Bjorkelund and Nivre, 2015)
- reimplementation of the head-selection parser of Zhang et al. (2017) (HSEL).

Experiments

		LAS			CNC		
		IMS	RBG	HSEL	IMS	RBG	HSEL
germanic	de	**84.3**	83.8	82.0	**79.7**	78.9	77.1
	en	**86.4**	86.3	86.0	**82.8**	82.2	82.3
iranian	fa	83.4	83.1	**83.9**	80.5	79.5	**80.8**
romance	ca	**89.5**	88.8	89.1	**84.0**	82.7	83.6
	es	**85.6**	85.2	85.2	**78.6**	77.5	78.0
	fr	**85.6**	84.4	85.2	**79.4**	77.6	78.6
	it	**89.6**	88.8	89.3	**84.3**	82.9	83.9
	ro	**79.9**	79.6	78.6	**75.4**	74.6	73.3
slavic	bg	**86.9**	84.9	85.6	**83.7**	80.8	81.7
	cs	**87.8**	86.1	85.7	**86.1**	83.9	83.5
	hr	79.9	**80.7**	78.1	77.2	**77.6**	74.9
	ru	**89.5**	**89.5**	86.8	**88.0**	87.8	84.4
sinitic	zh	**81.8**	79.4	80.4	**80.6**	77.9	79.1
finnic	et	**84.1**	83.9	75.3	**83.0**	82.6	73.0
turkic	tr	73.5	**75.1**	62.5	71.9	**73.4**	59.1

Table 2: LAS (excluding punctuation) and CNC (content dependencies only) on the test sets of the original treebanks.

Experiments

	lang	IMS CNC	IMS Δ	RBG CNC	RBG Δ	HSEL CNC	HSEL Δ
ger	de	81.0	1.3	81.2	2.3	78.0	0.9
	en	83.6	0.8	83.4	1.2	83.6	1.3
ira	fa	84.2	3.7	83.4	3.9	83.6	2.8
rom	ca	85.6	1.6	85.0	2.3	84.9	1.3
	es	80.5	1.9	80.8	3.3	79.9	1.9
	fr	81.9	2.5	80.7	3.1	80.4	1.8
	it	86.1	1.8	86.1	3.2	85.5	1.6
	ro	75.7	0.3	75.3	0.7	73.6	0.3
sla	bg	85.4	1.7	83.8	3.0	83.8	2.1
	cs	87.3	1.2	85.2	1.3	84.2	0.7
	hr	77.4	0.2	77.3	-0.3	73.2	-1.7
	ru	89.2	1.2	88.7	0.9	82.1	-2.3
sin	zh	81.9	1.3	78.9	1.0	79.2	0.1
fin	et	84.4	1.4	82.8	0.2	74.7	1.7
tur	tr	71.6	-0.3	71.8	-1.6	58.3	-0.8

Table 3: CNC for the converted treebanks and differences Δ to the CNC obtained on the original treebanks.

metric	orig	cop	prep	coord	c-p-c	Δ
			Turkish			
with punc	77.4	76.9	76.6	76.7	76.4	-1.0
w/o punc	75.1	74.4	74.1	74.2	73.8	-1.3
CNC	73.4	72.9	72.6	71.9	71.8	-1.6
core	65.9	65.3	65.9	64.7	**67.1**	+1.2
non-core	75.5	74.9	74.4	73.9	73.2	-2.3
func	85.6	84.2	83.4	**88.2**	**86.0**	+0.4
			Croatian			
with punc	80.2	78.7	79.4	**81.0**	80.1	-0.1
w/o punc	80.7	79.0	80.0	**81.5**	80.5	-0.2
CNC	77.7	75.5	76.9	**78.6**	77.3	-0.4
core	81.1	**81.5**	81.0	**81.7**	**81.9**	+0.7
non-core	76.8	74.0	75.9	**77.8**	76.1	-0.9
func	88.5	87.9	87.9	**89.1**	**88.7**	+0.2

Table 4: Results for different label sets for Turkish and Croatian (RBG parser) and difference (Δ) between original and converted treebank (**cop-prep-coord**).

Experiments

	Lang	orig	cop	prep	coord	c-p-c
ger	de	3.4	0.98	1.01	1.03	1.03
	en	2.9	1.00	1.04	1.03	1.07
ira	fa	3.5	0.97	0.99	1.02	0.97
rom	ca	3.1	1.00	1.06	1.03	1.09
	es	2.8	0.99	1.07	1.04	1.11
	fr	2.8	0.99	1.07	1.03	1.09
	it	2.7	1.00	1.05	1.02	1.08
	ro	2.7	1.00	1.04	1.04	1.07
sla	bg	2.5	1.01	1.05	1.02	1.08
	cs	2.8	1.00	1.58	1.03	1.06
	hr	2.8	1.00	1.03	1.04	1.08
	ru	2.7	1.00	1.02	1.03	1.05
sin	zh	3.6	1.00	0.98	1.01	1.00
fin	et	2.6	1.00	1.00	1.03	1.02
tur	tr	2.6	1.00	1.01	1.01	1.02

Table 5: Avg. dependency length in the original treebank and DLM ratio for each modification

$$DLMRatio = \sum_s \frac{DL_s}{|s|^2} / \sum_s \frac{ModDL_s}{|s|^2} \quad (1)$$

Experiments

	lang	Δ cop	Δ prep	Δ coord	Δ c-p-c
ger	de	-0.26	-0.03	0.03	-0.23
	en	-0.56	-0.19	-0.01	-0.72
ira	fa	-0.73	0.07	0.02	-0.60
rom	ca	0.09	0.07	-0.01	0.16
	es	-0.19	-0.19	0.02	-0.36
	fr	-0.16	-0.15	0.04	-0.27
	it	-0.22	-0.11	0.02	-0.29
	ro	-0.13	0.17	0.04	0.09
sla	bg	-0.31	-0.10	0.05	-0.34
	cs	-0.30	0.20	0.07	0.03
	hr	0.16	0.21	0.03	0.41
	ru	0.17	0.19	0.05	0.41
sin	zh	-0.25	-0.00	0.03	-0.19
fin	et	-0.37	0.16	0.04	-0.16
tur	tr	0.19	0.28	0.03	0.50

Table 6: Difference (Δ) between avg. unlexicalised arc direction entropy (ADE) in the original treebank and in the modified treebanks

$$H(Dir|Rel, H, D) = \sum_{rel,h,d} p(rel, h, d) H(Dir|rel, h, d)$$