Should Have, Would Have, Could Have

Investigating Verb Group Representations for Parsing with Universal Dependencies

UPPSALA UNIVERSITET

Miryam de Lhoneux and Joakim Nivre

Department of Linguistics and Philology
Uppsala University
{miryam.de_lhoneux,joakim.nivre}@lingfil.uu.se

UPPSALA UNIVERSITET

1. Introduction

- problem: UD is believed to be suboptimal for parsing
- solution: Create a parsing representation (de Marneffe et al., 2014)
- focus of the study: verb groups

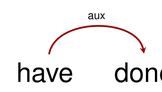

Figure 1: *MS verb group: the auxiliary is the head*

Figure 2: *UD/PDT verb group : the main verb is the head*

UD uses PDT style but MS is better for parsing

(Nilsson et al., 2006, 2007; Schwartz et al., 2012)

2.3 Data

Treebank	#S	#W	%A
SDT	1,936	35K	9.45
PDT	80,407	1,382K	1.38
Basque	7,194	97K	8.51
Bulgarian	10,022	141K	1.03
Croatian	3,757	84K	3.87
Czech	77,765	1,333K	0.92
Danish	5,190	95K	2.29
English	14,545	230K	2.85
Estonian	1,184	9K	0.73
Finnish	12,933	172K	1.49
Finnish-FTB	16,913	143K	2.89
French	16,148	394K	1.45
German	14,917	282K	1.05
Greek	2,170	53K	0.36
Hebrew	5,725	147K	0.15
Hindi	14,963	316K	3.27
Italian	12,188	260K	1.87
Norwegian	18,106	281K	2.60
Old_Church_Slavonic	5,782	52K	0.35
Persian	5,397	137K	1.40
Polish	7,500	76K	0.97
Portuguese	9,071	207K	0.20
Romanian	557	11K	2.88
Slovenian	7,206	126K	4.57
Spanish	15,739	424K	0.89
Swedish	4,807	76K	2.37
Tamil	480	8K	5.30

Table 1: *Stats on train + dev; S= sentence, W=word; A=aux dependencies.*

2.5 Software

- Parser: MaltParser (Nivre et al., 2006) with default settings and UD (coarse) PoS tags.
- Transformation algorithms: released as part of oDETTE (DEpendency Treebank Transformation and Evaluation). https://github.com/mdelhoneux/oDETTE

3.2 Error analysis

The baseline consistently outperforms the transformed model on the punctuation dependency relation. Punctuation is most often attached to the main verb. The transformed model is bad at identifying the main verb.

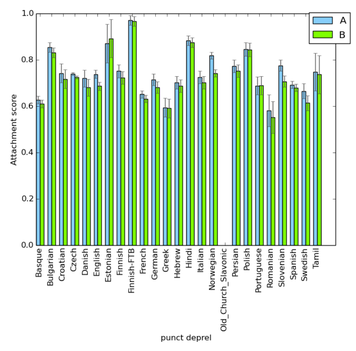

Figure 3: *F1 score and error margin in parsed test set*

2.1 Transformation Algorithm: UD to MS

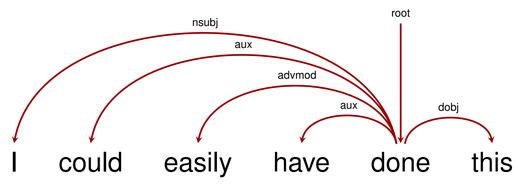

Figure 4: *UD sentence with a VG*

Algorithm:

1. Find main verb and collect auxiliaries set
2. Head of main verb becomes head of outermost auxiliary
3. Make a chain from outermost auxiliary to main verb

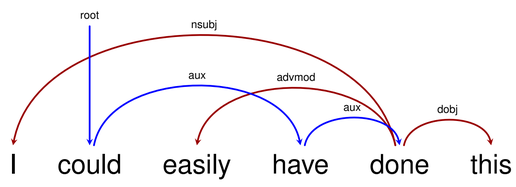

Figure 5: *Intermediate representation*

2.4 Pipeline

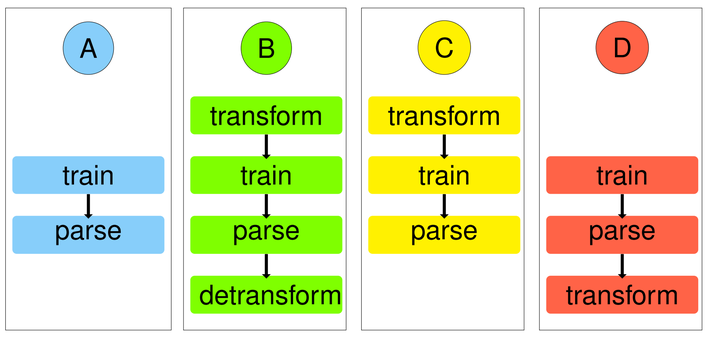

Figure 7: *Pipeline*

Model/Gold	UD	MS
UD	A	B
MS	D	C

Table 2: *Summary of Figure 7*

3.3 Role of POS tags ambiguity

Were improvements in PDT and SDT the result of POS disambiguation?

POS	main verb	aux
Verb-main	72.81	0.22
Verb-copula	22.30	95.95

Table 3: *2 main verb group POS tags in SDT*

We modify POS tags to create 3 treebanks:

- τ_o : original treebank
- τ_d : disambiguated treebank
- τ_a : ambiguous treebank

	A	B	Δ
SDT τ_d	**67.8**	67.4	-0.4
SDT τ_o	65.7	**66.2**	0.5
SDT τ_a	64.2	**65.4***	1.2
PDT τ_d	69.2	69.2	0.0
PDT τ_o	68.5	**68.8****	0.3
PDT τ_a	68.2	**68.4***	0.2

Table 4: *LAS on (A) and (B) with different levels of POS tag ambiguity. Δ = (B) - (A)*

✓The hypothesis seems to hold for SDT.
Less clear for PDT, maybe due to the use of predicted POS tags in experiments.

3.4 Predicted vs gold POS tags

Can UD benefit from the transformation when using predicted POS tags?
✗ It seems not.

POS tag	A	B	Δ
gold	76.8	75.7**	-1.1
predicted	76.4	75.6**	-0.8

Table 5: *LAS on UD_Swedish. Δ = (B) - (A)*

2. Methodology

4. Reattach main verb dependents according to their position compared to the verb group

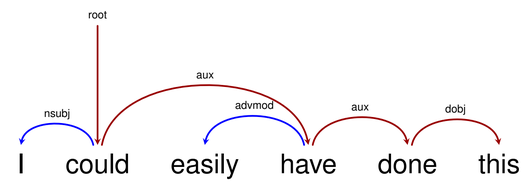

Figure 6: *MS representation*

2.2 Back Transformation: MS to UD

1. Find main verb and collect auxiliaries set
2. Attach auxiliaries to main verb
3. Attach auxiliaries dependents to main verb

We obtain 100% back transformation accuracy on all but 4 treebanks.

3. Results

3.1 Effect of VG Transformation on Parsing

UD language	A	B	C	D
Basque	**64.4**	63.8**	64.0	64.4
Bulgarian	**83.4**	83.2*	82.5	82.9
Croatian	**75.9**	74.6**	73.7	75.9
Czech	**80**	76.5**	76.4	79.9
Danish	**75.9**	75.2**	74.8	75.8
English	**81.7**	80.4**	80.2	81.5
Estonian	77.1	**77.8**	77.6	77.0
Finnish	**66.9**	66.4*	65.9	66.4
Finnish-FTB	**71.3**	70.4**	72.1	72.5
French	**82.1**	81.6**	81.3	81.8
German	**76.6**	76.0**	75.4	76.1
Greek	75.2	**75.3**	75.1	75.2
Hebrew	**78.4**	77.9**	77.9	78.5
Hindi	**85.4**	84.2**	84.9	85.2
Italian	**83.8**	83.6	83.3	83.6
Norwegian	**84.5**	82.0**	81.7	84.5
Old_Church_Slavonic	**68.8**	68.7	68.7	68.9
Persian	**81.1**	79.8**	79.8	81.1
Polish	**79.4**	79.1	79.0	79.3
Portuguese	81.3	**81.5**	81.6	81.3
Romanian	64.2	62.5*	64.0	64.6
Slovenian	**80.8**	79.7**	79.8	80.8
Spanish	**81.5**	81.2**	81.2	81.4
Swedish	**76.8**	75.7**	75.6	76.7
Tamil	**67.2**	67.1	67.4	67.5

Table 6: *LAS with the 4 versions of the treebank.*

MS is better than UD for parsing	B > A ✗
MS is easier to learn than UD	C > A ✗
Symmetry in differences	A - B = C - D ✗

Table 7: *Hypotheses*

4. Conclusion

- Verb groups should stay as is in UD.
- Gains from transforming from PDT style to MS style in previous studies were probably obtained because the approach helped disambiguate POS tags.

Future work

- Looking at other parsing models.
- More in-depth error analysis.
- Looking at other representations (e.g. PPs).

References

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014), Reykjavik, Iceland, May 26-31, 2014.*. pages 4585–4592.

Jens Nilsson, Joakim Nivre, and Johan Hall. 2006. Graph transformations in data-driven dependency parsing. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL-44, pages 257–264.

Jens Nilsson, Joakim Nivre, and Johan Hall. 2007. Generalizing tree transformations for inductive dependency parsing. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. pages 2216–2219.

Roy Schwartz, Omri Abend, and Ari Rappoport. 2012. Learnability-based syntactic annotation design. In *COLING*. volume 24, pages 2405–2422.