The parse is darc and full of errors:
Universal dependency parsing with transition-based and graph-based algorithms

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Kuan Yu ⟨kuan.yu@student.uni-tuebingen.de⟩
Pavel Sofroniev ⟨pavel.sofroniev@student.uni-tuebingen.de⟩
Erik Schill ⟨erik.schill@student.uni-tuebingen.de⟩
Erhard Hinrichs ⟨erhard.hinrichs@uni-tuebingen.de⟩

Department of Linguistics, University of Tübingen

In brief

- two simple systems for dependency parsing using neural networks
- **darc**: a transition-based non-projective/projective parser
  - the official system, ranked 12th among 33 systems
  - simpler and more restricted than UDPipe's Parsito (Straka, Hajič, Straková, and Hajič jr. 2015)
  - achieved comparable results with baseline & ÚFAL
- **mstnn**: a graph-based non-projective unlabeled parser
  - with a standalone labeler

	darc	mstnn	baseline	ÚFAL
all treebanks	68.41	61.13	68.35	69.52
big treebanks	73.31	65.84	73.04	74.38
small treebanks	52.46	48.40	51.80	53.75
parallel test-sets	67.96	60.49	68.33	69.00
surprise languages	34.47	24.04	37.07	35.96

Neural network learner

- same structure for both parsers
- fully connected feedforward network
- AdaMax optimization
- 2 layers of 256 hidden ReLU
- 25% dropout on hidden layers
- unit-norm constraint on embeddings
- pretrained FORM & LEMMA embeddings

Machine learning features

	min	max	avg	dim
FORM	70	58 562	9 070	32
LEMMA	89	29 972	6 321	32
UPOSTAG	13	19	18	12
DEPREL	21	55	37	16

trainable embedding shapes

FEATS		
count of	max	avg
types	2 487	430
hapaxes	561	92
unique entries	112	44

morphological entries

darc

- arc-standard system plus swap (Nivre 2008; Nivre 2009)
- static oracle, lazy with swap, greedy decoding
- 18 graph nodes as inputs (Chen and Manning 2014)
  - top 3 words on stack & buffer
  - 1st & 2nd leftmost & rightmost children of the top two stack nodes
  - leftmost-of-leftmost & rightmost-of-rightmost children of the top two stack nodes

mstnn

- neural network augmented maximum spanning tree parser (McDonald et al. 2005)
- sigmoid probabilities as weight scores
- first order inputs
  - the two nodes and their left & right
  - the left-of-left & right-of-right neighbors
  - distance between the two nodes

Treatments for datasets

- only used *Universal Dependencies version 2.0* (Nivre et al. 2017) treebanks for model training
- new UDPipe (Straka, Hajič, and Straková 2016) models with baseline hyperparameters for segmentation and tagging
- for big treebanks: trained on train-sets & tuned on dev-sets
- for small treebanks: consulted the baseline splits of train-, tune-, & dev-sets
- for parallel test-sets: picked models of the same languages
- for surprise languages: trained delexicalized models on the closest treebanks based on the scores on the sample data

  Buryat: *Polish*, Kurmanji: *Polish*, North Sami: *Finnish*, Upper Sorbian: *Slovenian*

Summary

- openly available at `https://github.com/CoNLL-UD-2017/darc`
- powerful machine learning tools may alleviate the drawbacks of transition-based parsing
- carefully constructed input features help
- ultimately, an integrated approach is more desirable than a pipeline

LAS

	darc	mstnn
Ancient Greek	**58.20**	54.78
Ancient Greek PROIEL	66.21	62.81
Arabic	65.49	60.61
Arabic PUD	44.10	39.86
Basque	68.08	64.15
Bulgarian	84.51	75.19
Buryat	15.61	19.69
Catalan	85.39	73.91
Chinese	56.44	47.07
Croatian	76.96	68.17
Czech	81.92	73.89
Czech PUD	79.54	71.37
Czech CAC	81.78	73.78
Czech CLTT	73.57	67.26
Danish	73.67	64.55
Dutch	68.94	60.85
Dutch LassySmall	79.89	71.17
English	75.83	65.25
English PUD	77.67	64.70
English LinES	72.98	63.62
English ParTUT	74.39	63.91
Estonian	59.75	54.62
Finnish	74.93	67.19
Finnish PUD	78.49	70.14
Finnish FTB	**75.43**	70.93
French	**80.50**	69.90
French PUD	73.06	64.91
French ParTUT	78.84	70.85
French Sequoia	79.44	71.00
Galician	77.17	69.97
Galician TreeGal	65.19	59.76
German	68.02	63.77
German PUD	65.09	60.76
Gothic	61.92	58.54
Greek	79.05	73.53
Hebrew	57.13	50.84
Hindi	**87.50**	80.99
Hindi PUD	50.98	47.79
Hungarian	65.17	60.05
Indonesian	73.58	63.38
Irish	**62.97**	57.55
Italian	85.04	76.45
Italian PUD	83.79	73.43
Japanese	72.88	64.27
Japanese PUD	75.69	66.85
Kazakh	23.68	22.28
Korean	58.30	56.32
Kurmanji	**33.06**	23.54
Latin	45.29	43.53
Latin ITTB	**76.22**	69.52
Latin PROIEL	59.52	56.23
Latvian	**62.03**	54.60
North Sami	**34.89**	21.67
Norwegian Bokmaal	82.29	71.53
Norwegian Nynorsk	80.99	69.14
Old Church Slavonic	**66.37**	63.87
Persian	77.59	66.59
Polish	79.72	77.06
Portuguese	81.40	72.16
Portuguese PUD	73.65	64.68
Portuguese BR	84.98	72.49
Romanian	**80.42**	70.10
Russian	74.83	68.63
Russian PUD	68.61	63.11
Russian SynTagRus	86.39	79.06
Slovak	73.49	68.29
Slovenian	81.05	73.21
Slovenian SST	47.41	42.52
Spanish	81.27	69.53
Spanish PUD	77.49	66.78
Spanish AnCora	84.06	72.37
Swedish	76.45	65.34
Swedish PUD	68.94	59.68
Swedish LinES	73.62	64.12
Turkish	**54.70**	52.44
Turkish PUD	**34.37**	32.84
Ukrainian	**62.03**	56.39
Upper Sorbian	54.30	31.24
Urdu	**77.21**	70.39
Uyghur	**34.28**	**34.32**
Vietnamese	37.31	31.82

score ≥ `UDPipe 1.1 (baseline)`
score ≥ `UDPipe 1.2 (ÚFAL)`

References

Chen, Danqi and Christopher D Manning (2014). "A Fast and Accurate Dependency Parser using Neural Networks." In: *EMNLP*, pp. 740–750.
McDonald, Ryan et al. (2005). "Non-projective dependency parsing using spanning tree algorithms". In: *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 523–530.
Nivre, Joakim (2008). "Algorithms for deterministic incremental dependency parsing". In: *Computational Linguistics* 34.4, pp. 513–553.
— (2009). "Non-projective dependency parsing in expected linear time". In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*. Association for Computational Linguistics, pp. 351–359.
Nivre, Joakim et al. (2017). *Universal Dependencies 2.0*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, Prague, `http://hdl.handle.net/11234/1-1983`. URL: `http://hdl.handle.net/11234/1-1983`.
Straka, Milan, Jan Hajič, and Jana Straková (2016). "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing". In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association. ISBN: 978-2-9517408-9-1.
Straka, Milan, Jan Hajič, Jana Straková, and Jan Hajič jr. (2015). "Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle". In: *Proceedings of Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14)*.