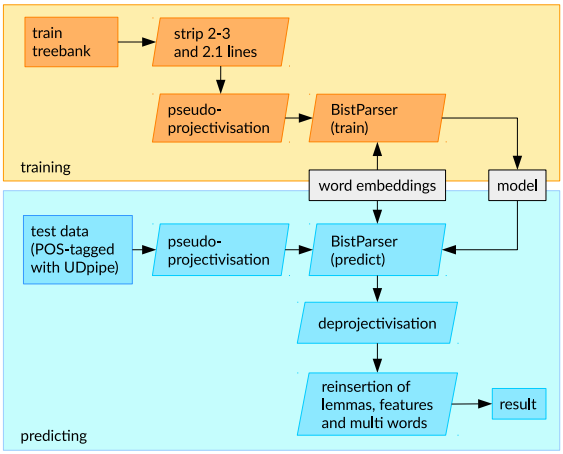
CoNLL 2017 UD Shared Task

Multi-Model and Crosslingual Dependency Analysis

Johannes Heinecke, Munshi Asadullah
Orange Labs, Lannion, France

Architecture



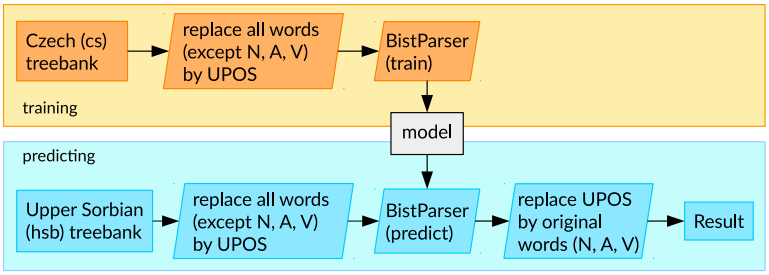
BistParser modifications: One depencency tree per sentence

Training:

- hidden layer size 40, 50 or 100, depending on language
- other BistParser options used: `--k 3 --lstmdims 125 --lstmlayers 2 --bibi-lstm --usehead --userl`
- word embeddings for all languages (except Gothic)
  - all words in lowercase (if applicable)
  - punctuation separated from words
  - `word2vec` standard options except `-size {300,500}` and `-window 10`

Surprise Languages

Two crosslingual approaches: training (1) on a mix of 23 languages and (2) on a typologically close language (*hsb* → *cs*, *sme* → *fi*, *kmr* → *fa*, *bxr* → *hi*), both without word embeddings: (2) gave much better results.



Example of modified CONLL (cols. 1, 2 and 4) used for training (i.e. *cs*, shown below left) and prediction (in this case *hsb*, below right):

training data (*cs*)			test data (*hsb*)		
1	*Manažeři*	NOUN	1	Njejsu	VERB
2	*rozhodují*	VERB	2	DET	DET
3	ADV	ADV	3	archeologiske	ADJ
4	ADP	ADP	4	dokłady	NOUN
5	*místě*	NOUN	5	ADP	ADP
6	PUNCT	PUNCT	...		

23 language mix		Upper Sorbian (hsb)		Northern Sami (sme)		Kurmanji (kmr)		Buryat (bxr)	
Upper Sorbian (100[1])	63.2%	**cs (100)**	**69.5%**	**fi (100)**	**52.9%**	**fa (100)**	**36.7%**	**hi (50)**	**32.0%**
Northern Sami (50)	49.2%	cs (50)	67.5%	fiu[2] (100)	51.7%	fa (50)	35.8%	ur (50)	28.0%
Kurmanji (100)	29.2%	pl (50)	56.9%	fiu (50)	49.7%	hi (50)	22.2%	tr (100)	27.6%
Buryat (50)	26.3%	pl (100)	51.9%	fi (50)	50.8%	ur (50)	20.6%	fi (100)	21.8%
								ja (50)	18.0%

[1] Number indicates hidden layer size. [2] Mix of Fenno-Ugric languages, here *fi*, *et* and *hu*.

Results

- **10th position with LAS 68.61% (improved to 69.75% after bug fixes)**
- 9th position with Content Word LAS (CLAS) as evaluation metrics: 64.15%
- 8th position on surprise languages: 38.72% (7th position with CLAS: 34.28%)

Runtime (on Tira VM, Ubuntu Xenial): 3 hours (all treebanks), using <16GB

Contact: **johannes.heinecke@orange.com**