UPPSALA UNIVERSITET

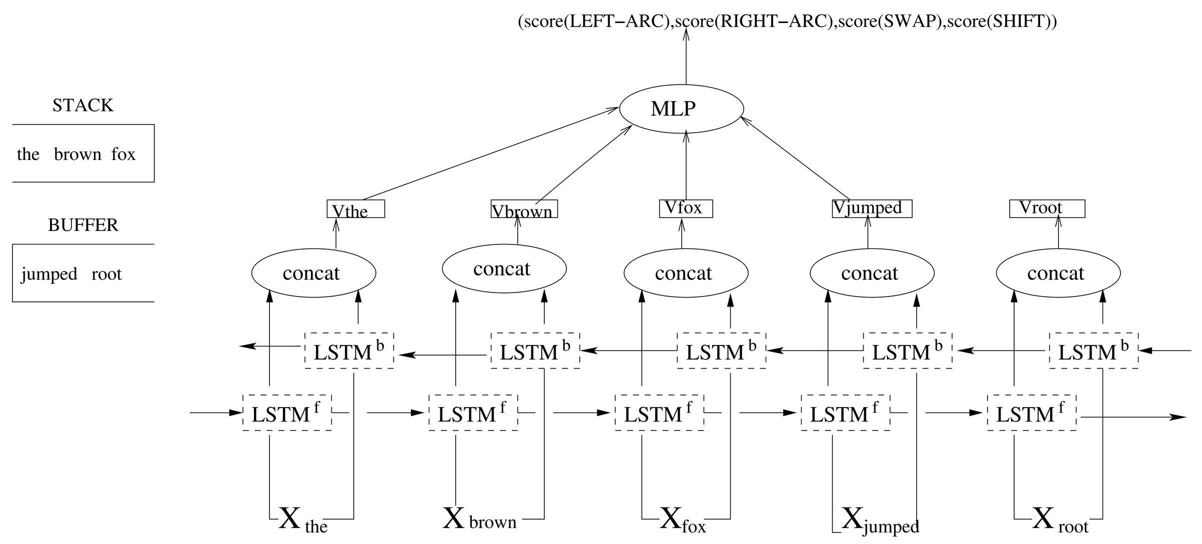
Parser Training with Heterogeneous Treebanks

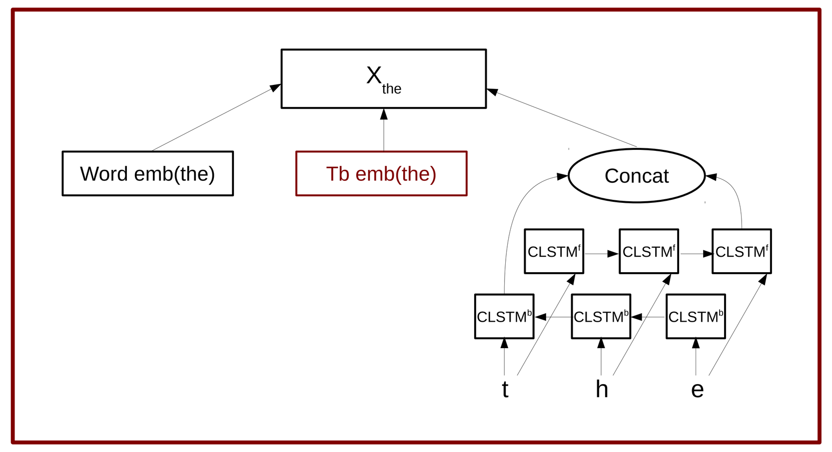
Sara Stymne, Miryam de Lhoneux, Aaron Smith and Joakim Nivre

Introduction

- **Problem:** How can we improve parsing when there are several, potentially heterogeneous treebanks for a language?
- Treebank diversity
  - Annotation scheme
  - Language variant
  - Spoken/written language
  - Genres and domains
  - Treebank size
  - Annotation quality and consistency
- **This work:**
  - Investigate previously proposed strategies
  - Introduce treebank embeddings

System Architecture





Dims word emb	100
Dims char emb	12
Dims treebank emb	12
Word LSTM dims	125
Char LSTM dims	50
LSTM dropout	0.33
Word dropout (α)	0.25
Epochs	30
Epochs fine-tuning	10

Strategies

Single
One model per treebank
+ Simple
– Does not take advantage of all data
– Separate models for each treebank

Concatenation
One model per language, on concatenated data
+ Simple
– Does not take treebank differences into account
+ A single model per language

Concatenation + fine-tuning
Fine-tune a different model for each treebank, based on the concatenation (Che et al., 2017, Shi et al., 2017)
– Needs more training than previous models
– Separate models for each treebank
+ Takes treebank differences into account

Treebank embeddings
Train a single model per language, but use a treebank embedding to represent the treebank each word comes from. Similar to language embeddings (Ammar et al., 2016)
+ Simple
+ Takes treebank differences into account
+ A single model per language

Other approaches (not in this paper)
- 1-hot treebank representation: similar to our approach, but with 1-hot representation rather than embedding (Lim & Poibeau, 2017).
- Adversarial learning: combine treebank specific models with a joint model where treebank identification is an adversarial task (Sato et al., 2017). Effective, especially on small treebanks, but more complicated than our model.

Parsing Unseen Data

When parsing unseen data, we need to choose an existing treebank: **proxy treebank**
- Single: the treebank used to train a model
- Concatenation: N/A
- Concatenation + fine-tuning: the treebank used for fine-tuning
- Treebank embeddings: the treebank embedding to use in the model

Experiments

- Universal dependencies version 2.1
- Standardized annotation scheme, but still many differences
- 9 languages:
  - at least 2 training treebanks
  - test set without training data (PUD)

Results

			Same treebank test set				PUD (unseen) test set			
Language	**Treebank**	**Size**	SINGLE	CONCAT	C+FT	TB-EMB	SINGLE	CONCAT	C+FT	TB-EMB
Czech	PDT	68495	86.7	87.5^{+}	**88.3***	87.2^{+}	**81.7**	**81.7**	81.6	81.2
	CAC	23478	86.0	87.8^{+}	88.1^{+}	**88.5**$^{+}$	75.0		81.3	81.1
	FicTree	10160	84.3	89.3^{+}	**89.5**$^{+}$	89.2^{+}	66.1		79.8	80.3
	CLTT	860	72.5	86.2^{+}	**86.9**$^{+}$	86.0^{+}	42.1		80.8	80.9
English	EWT	12543	82.2	82.1	82.5	**83.0**	80.7	80.0	81.7^{*}	**81.9***
	LinES	2738	72.1	76.7^{+}	**77.3**$^{+}$	**77.3**$^{+}$	62.6		75.9	74.5
	ParTUT	1781	80.5	83.5^{+}	85.4^{+}	**85.7**$^{+}$	68.0		78.1	76.9
Finnish	FTB	14981	$76.4^{\times}$	74.4	80.1^{*}	**80.6***	46.7	73.0	54.6	53.1
	TDT	12217	$78.1^{\times}$	70.6	**80.6***	80.3^{*}	$78.6^{\times}$		**81.3***	80.9^{*}
French	FTB	14759	83.2	83.2	83.9^{*}	**84.1***	72.0	79.4	76.7	74.1
	GSD	14554	84.5	84.1	85.3	**85.6**$^{\times}$	79.1		80.2^{*}	**80.3***
	Sequoia	2231	84.0	86.0^{+}	**89.8***	89.1^{*}	69.5		78.1	77.6
	ParTUT	803	79.8	80.5	89.1^{*}	**90.3***	63.4		78.8	77.5
Italian	ISDT	12838	87.7	**87.9**	87.7	87.6	85.4	86.0	85.7	86.0
	PoSTWITA	2808	71.4	76.7^{+}	76.8^{+}	**77.0**$^{+}$	68.5		85.7	85.3
	ParTUT	1781	83.4	89.2^{+}	**89.3**$^{+}$	88.8^{+}	77.4		85.8^{+}	**86.1**$^{+}$
Portuguese	GSD	9664	88.3	87.3	89.0^{*}	**89.1***	74.0	76.8^{+}	75.2	74.9
	Bosque	8331	84.7	84.2	$86.2^{\times}$	**86.3***	75.2		77.5^{+}	**77.6**$^{+}$
Russian	SynTagRus	48814	$90.2^{\times}$	89.4	**90.4**$^{\times}$	**90.4**$^{\times}$	66.0	68.7	66.3	66.4
	GSD	3850	$74.7^{\times}$	73.4	79.8^{*}	**80.8***	$70.1^{\times}$		77.6^{*}	**78.0***
Spanish	AnCora	14305	$87.2^{\times}$	86.2	$87.5^{\times}$	**87.6**$^{\times}$	75.2	79.9	77.7	76.4
	GSD	14187	84.7	83.0	$85.8^{\times}$	**86.2***	79.8		80.8^{+}	**80.9***
Swedish	Talbanken	4303	79.6	79.1	80.2	**80.6**$^{\times}$	70.3	72.0^{+}	73.2^{*}	**73.6***
	LinES	2738	74.3	76.8	**77.3**$^{+}$	77.1^{+}	64.0		70.0	69.0
Average			81.4	82.7^{+}	**84.9***	**84.9***	77.9	77.5	80.0^{*}	**80.1***

$^{+}$ significantly better than SINGLE $^{\times}$ significantly better than CONCAT * significantly better than SINGLE+CONCAT

Conclusion

- Combining treebanks is beneficial, especially for small treebanks
- Treebank embeddings successful
  - At least on par with other methods
  - Simple model
  - Works for many different scenarios
- Choice of proxy treebank very important

References

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah Smith. 2016. Many languages, one parser. TACL, 4:431444.
Wanxiang Che, Jiang Guo, Yuxuan Wang, Bo Zheng, Huaipeng Zhao, Yang Liu, Dechuan Teng, and Ting Liu. 2017. The HIT-SCIR system for end-to-end parsing of universal dependencies. CoNLL 2017.
KyungTae Lim and Thierry Poibeau. 2017. A system for multilingual dependency parsing based on bidirectional LSTM feature representations. CoNLL 2017.
Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. 2017. Adversarial training for cross-domain universal dependency parsing. CoNLL 2017.
Tianze Shi, Felix G. Wu, Xilun Chen, and Yao Cheng. 2017. Combining global models for parsing universal dependencies. CoNLL 2017.