

EPE 2017

Building an Infrastructure for Extrinsic Parser Evaluation

Stephan Oepen

Jari Björne, Filip Ginter, Richard Johansson, Emanuele Lapponi, Joakim Nivre, Anders Søgaard, Erik Velldal, Lilja Øvrelid

epe-organizers@nlpl.eu

Some Near-Authentic Quotes and Reflections

To me, the ultimate goal of our new field of Computational Linguistics is to build machines that, in a suitable interpretation of that term, 'understand' human language.

(Martin Kay, 1960s)



Some Near-Authentic Quotes and Reflections

To me, the ultimate goal of our new field of Computational Linguistics is to build machines that, in a suitable interpretation of that term, 'understand' human language.

(Martin Kay, 1960s)

20 Years of Progress in Statistical Parsing

- Parsing into PTB-style trees has been a crisp task for many years;
- great advances: representations, algorithms, probabilistic models;
- F_1 : 0.84 (Magerman, 1994) \rightarrow 0.91 (Charniak & Johnson, 2005);
- some ten years later, neural advances: 93.8 (Choe & Charniak, 2016).



DM: DELPH-IN MRS (Bi-Lexical) Dependencies

- DeepBank: Fresh HPSG-style annotation, including logical-form semantics;
- 'lossy' reduction of MRS meaning representations to bi-lexical dependencies.



DM: DELPH-IN MRS (Bi-Lexical) Dependencies

- DeepBank: Fresh HPSG-style annotation, including logical-form semantics;
- 'lossy' reduction of MRS meaning representations to bi-lexical dependencies.

PAS: Enju Predicate—Argument Structures

- Enju Treebank: Projection of (complete) PTB syntax to HPSG derivations;
- semantic analyses take form of lexicalized predicate—argument structures.



DM: DELPH-IN MRS (Bi-Lexical) Dependencies

- DeepBank: Fresh HPSG-style annotation, including logical-form semantics;
- 'lossy' reduction of MRS meaning representations to bi-lexical dependencies.

PAS: Enju Predicate—Argument Structures

- Enju Treebank: Projection of (complete) PTB syntax to HPSG derivations;
- semantic analyses take form of lexicalized predicate—argument structures.

PSD: Parts of the Prague Tectogrammatical Layer

- Include all nodes from Prague t-trees that correspond to surface tokens;
- re-attach functors of generated nodes; project dependencies to conjuncts.



DM: DELPH-IN MRS (Bi-Lexical) Dependencies

- DeepBank: Fresh HPSG-style annotation, including logical-form semantics;
- 'lossy' reduction of MRS meaning representations to bi-lexical dependencies.

PAS: Enju Predicate—Argument Structures

- Enju Treebank: Projection of (complete) PTB syntax to HPSG derivations;
- semantic analyses take form of lexicalized predicate—argument structures.

PSD: Parts of the Prague Tectogrammatical Layer

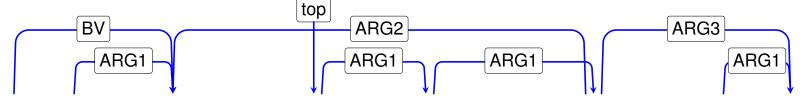
- Include all nodes from Prague t-trees that correspond to surface tokens;
- re-attach functors of generated nodes; project dependencies to conjuncts.

WSJ 00–20 for Training (802,717 Tokens); Section 21 for Testing (31,948).

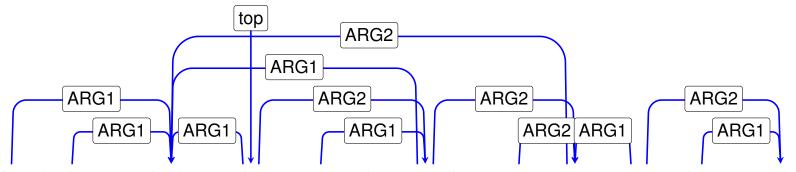


 $\sum_{i=1}^{N}$

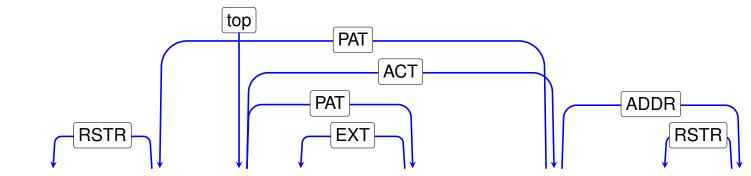
PAS



A similar technique is almost impossible to apply to other crops.



A similar technique is almost impossible to apply to other crops.

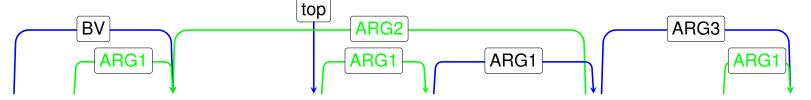


A similar technique is almost impossible to apply to other crops .

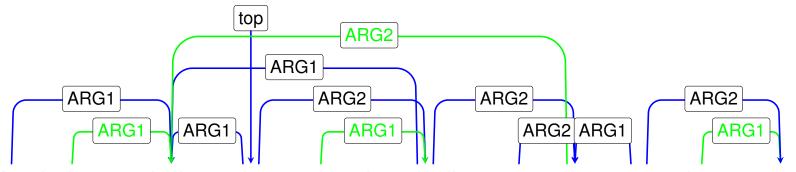


 \sum_{Ω}

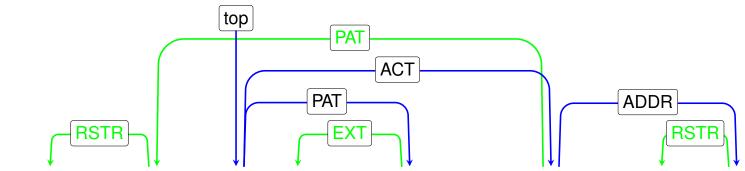
PAS



A similar technique is almost impossible to apply to other crops .



A similar technique is almost impossible to apply to other crops.

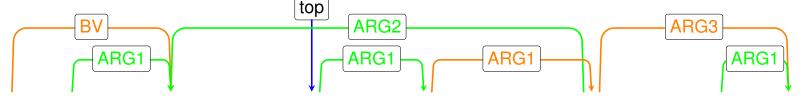


A similar technique is almost impossible to apply to other crops .

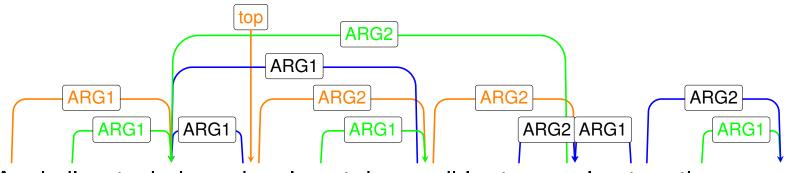


 \mathbb{Z}

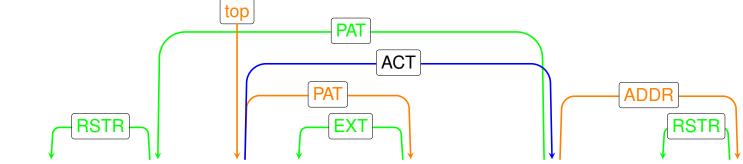
PAS



A similar technique is almost impossible to apply to other crops.



A similar technique is almost impossible to apply to other crops.



A similar technique is almost impossible to apply to other crops .



			D	M			P	AS			PS	SD	
	LF	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM
Peking	85.91	90.27	88.54	89.40	26.71	93.44	90.69	92.04	38.13	78.75	73.96	76.28	11.05
Priberam	85.24	88.82	87.35	88.08	22.40	91.95	89.92	90.93	32.64	78.80	74.70	76.70	09.42
Copenhagen- Malmö	80.77	84.78	84.04	84.41	20.33	87.69	88.37	88.03	10.16	71.15	68.65	69.88	08.01
Potsdam	77.34	79.36	79.34	79.35	07.57	88.15	81.60	84.75	06.53	69.68	66.25	67.92	05.19
Alpage	76.76	79.42	77.24	78.32	09.72	85.65	82.71	84.16	17.95	70.53	65.28	67.81	06.82
Linköping	72.20	78.54	78.05	78.29	06.08	76.16	75.55	75.85	01.19	60.66	64.35	62.45	04.01



			D	M			P	AS			PS	SD	
	LF	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM
Peking	85.91	90.27	88.54	89.40	26.71	93.44	90.69	92.04	38.13	78.75	73.96	76.28	11.05
Priberam	85.24	88.82	87.35	88.08	22.40	91.95	89.92	90.93	32.64	78.80	74.70	76.70	09.42
Copenhagen- Malmö	80.77	84.78	84.04	84.41	20.33	87.69	88.37	88.03	10.16	71.15	68.65	69.88	08.01
Potsdam	77.34	79.36	79.34	79.35	07.57	88.15	81.60	84.75	06.53	69.68	66.25	67.92	05.19
Alnade	76 76	79 42	77 24	78 32	NQ 72	25 65	ጸኃ 71	24 16	17 95	70 53	65 28	67 A1	ი6.82
				0	bser	vatio	ns						1.01

• Ensemble system (including graph parsers) best in 'closed' track;



			D	M			P	AS			PS	SD	
	LF	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM
Peking	85.91	90.27	88.54	89.40	26.71	93.44	90.69	92.04	38.13	78.75	73.96	76.28	11.05
Priberam	85.24	88.82	87.35	88.08	22.40	91.95	89.92	90.93	32.64	78.80	74.70	76.70	09.42
Copenhagen- Malmö	80.77	84.78	84.04	84.41	20.33	87.69	88.37	88.03	10.16	71.15	68.65	69.88	08.01
Potsdam	77.34	79.36	79.34	79.35	07.57	88.15	81.60	84.75	06.53	69.68	66.25	67.92	05.19
Δlnade	76 76	79 42	77 24	78 32	N9 72	25 65	୧୨ 71	24 16	17 95	70 53	65 2 <u>8</u>	67 2 1	ი6.82
				0	bser	vatio	ns						ŀ.01

- Ensemble system (including graph parsers) best in 'closed' track;
- high per-dependency scores: 76 92 F₁ for best 'closed' systems;



			D	M			P/	AS			PS	SD	
	LF	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM
Peking	85.91	90.27	88.54	89.40	26.71	93.44	90.69	92.04	38.13	78.75	73.96	76.28	11.05
Priberam	85.24	88.82	87.35	88.08	22.40	91.95	89.92	90.93	32.64	78.80	74.70	76.70	09.42
Copenhagen- Malmö	80.77	84.78	84.04	84.41	20.33	87.69	88.37	88.03	10.16	71.15	68.65	69.88	08.01
Potsdam	77.34	79.36	79.34	79.35	07.57	88.15	81.60	84.75	06.53	69.68	66.25	67.92	05.19
Δlnane	76 76	79 42	77 24	78 32	N9 72	25 65	Ջ 2 71	24 16	17 95	70 53	65 2 <u>8</u>	67 A1	ი6.82
				0	bser	vatio	ns						1.01

- Ensemble system (including graph parsers) best in 'closed' track;
- ◆ high per-dependency scores: 76 92 F₁ for best 'closed' systems;
- exact match sentence accuracy a bit less encouraging: 9 38 %;



			D	M			P/	AS			PS	SD	
	<u>LF</u>	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM
Peking	85.91	90.27	88.54	89.40	26.71	93.44	90.69	92.04	38.13	78.75	73.96	76.28	11.05
Priberam	85.24	88.82	87.35	88.08	22.40	91.95	89.92	90.93	32.64	78.80	74.70	76.70	09.42
Copenhagen- Malmö	80.77	84.78	84.04	84.41	20.33	87.69	88.37	88.03	10.16	71.15	68.65	69.88	08.01
Potsdam	77.34	79.36	79.34	79.35	07.57	88.15	81.60	84.75	06.53	69.68	66.25	67.92	05.19
ΔΙηρηρ	76 76	70 12	77 94	72 22	NQ 79	25 65	ჲე 71	24 16	17 05	70 53	65 2 <u>8</u>	67 2 1	ი6.82
				0	bser	vatio	ns						1.01

- Ensemble system (including graph parsers) best in 'closed' track;
- high per-dependency scores: 76−92 F₁ for best 'closed' systems;
- exact match sentence accuracy a bit less encouraging: 9 38 %;
- parsers based on (only) tree approximations not fully competitive;



			D	M			P/	AS			PS	SD	
	LF	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM
Peking	85.91	90.27	88.54	89.40	26.71	93.44	90.69	92.04	38.13	78.75	73.96	76.28	11.05
Priberam	85.24	88.82	87.35	88.08	22.40	91.95	89.92	90.93	32.64	78.80	74.70	76.70	09.42
Copenhagen- Malmö	80.77	84.78	84.04	84.41	20.33	87.69	88.37	88.03	10.16	71.15	68.65	69.88	08.01
Potsdam	77.34	79.36	79.34	79.35	07.57	88.15	81.60	84.75	06.53	69.68	66.25	67.92	05.19
Alnane	76 76	79 42	77 24	78 32	N9 72	25 65	Ջ 2 71	24 16	17 95	70 53	65 2 <u>8</u>	67 A1	ი6.82
				0	bser	vatio	ns						1.01

- Ensemble system (including graph parsers) best in 'closed' track;
- high per-dependency scores: 76 92 F₁ for best 'closed' systems;
- exact match sentence accuracy a bit less encouraging: 9 38 %;
- parsers based on (only) tree approximations not fully competitive;
- PAS overall easiest to parse, (labeling) PSD is noticeably harder;



			D	M			PA	AS		PSD				
	LF	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM	
Peking	85.91	90.27	88.54	89.40	26.71	93.44	90.69	92.04	38.13	78.75	73.96	76.28	11.05	
Priberam	85.24	88.82	87.35	88.08	3 22.40	91.95	89.92	90.93	32.64	78.80	74.70	76.70	09.42	
				(Comp	arisc	on						3.01	
• graph	adap	tation	of ('s	synta	ctic')	Turbo	Parso	er as	best '	open	ı' syst	em;	5.19 5.82 1.01	

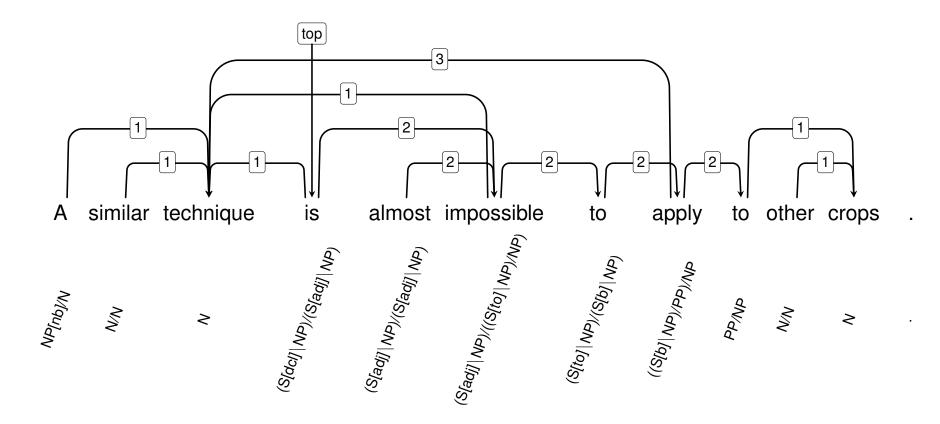
			D	M			PA	AS			PS	SD	
	ΙF	LP	LR	LF	LM	LP	LR	LF	LM	LP	LR	LF	LM
Priberam	86.27	90.23	88.11	89.16	26.85	92.56	90.97	91.76	37.83	80.14	75.79	77.90	10.68
CMU	82.42	84.46	83.48	83.97	08.75	90.78	88.51	89.63	26.04	76.81	70.72	73.64	07.12
Turku	80.49	80.94	82.14	81.53	08.23	87.33	87.76	87.54	17.21	72.42	72.37	72.40	06.82
Potsdam	78.60	81.32	80.91	81.11	09.05	89.41	82.61	85.88	07.49	70.35	67.33	68.80	05.42
Alpage	78.54	83.46	79.55	81.46	10.76	87.23	82.82	84.97	15.43	70.98	67.51	69.20	06.60
In-House	75.89	92.58	92.34	92.46	48.07	92.09	92.02	92.06	43.84	40.89	45.67	43.15	00.30



				D	M				PAS			Р	SD	
	Ī	F	LP	LR	LF	LIV	L LF	P LF	LF	LM	LP	LR	LF	LM
Peking	j 85	5.91 9	0.27 8	88.54	89.4	0 26.7	'1 93.4	44 90.6	992.0	4 38.1	3 78.7	5 73.96	6 76.28	3 11.05
Pribera	m 85	5.24 8	8.82 8	37.35	88.0	8 22.4	0 91.9	95 89.9	90.9	3 32.6	4 78.80	0 74.70	76.70	09.42
						Com	pari	son						3.01
- 0140	nh ad	lanta	ition (of ('s	synta	actic') Turl	ooPar	ser a	s best	t 'ope	n' sys	stem;	5.19 5.82
• gra	pri ac	αρισ		. (-							•			
		•		· ·			m se	veral	F_1 po	ints a	head	of the	field	01
		•		· ·			m se	veral	F ₁ po	ints a	head	of the	e field	01
		•		sten			m se	veral LR	F ₁ po	ints a	head	of the	field LF	01
) 'ln-l 	Hous	se' sy LR	rsten	ns p	erfor LM	LP	LR	LF	LM	LP	LR	LF	LM
• (ful) 'ln-l LF 86.27	Hous LP 90.23	se' sy LR 3 88.1	sten	ns p F	erfor LM 26.85	LP 92.56	LR 90.97	LF 91.76	LM	LP 80.14	LR 75.79	LF 77.90	LM 10.68
• (ful	LF 86.27 82.42	LP 90.23	LR 3 88.1 6 83.4	sten L 1 89	ns p .F .16 2	erfor LM 26.85	LP 92.56 90.78	LR 90.97 88.51	LF 91.76 89.63	LM 37.83	LP 80.14 76.81	LR 75.79 70.72	LF 77.90 73.64	LM 10.68 07.12
• (ful	LF 86.27 82.42 80.49	LP 90.23 84.44 80.94	LR 3 88.1 6 83.4 4 82.1	2 sten 1 89 8 83 4 81	ns p .F .16 2 .97 0	erfor LM 26.85 98.75	LP 92.56 90.78 87.33	LR 90.97 88.51 87.76	LF 91.76 89.63 87.54	LM 37.83 26.04	LP 80.14 76.81 72.42	LR 75.79 70.72 72.37	LF 77.90 73.64 72.40	LM 10.68 07.12 06.82
• (ful Priberam CMU Turku	LF 86.27 82.42 80.49 78.60	LP 90.23 84.44 80.94 81.33	LR 3 88.1 6 83.4 4 82.1 2 80.9	1 89 8 83 4 81	ns p .F .16 2 .97 0 .53 0	erfor LM 26.85 98.75 98.23	LP 92.56 90.78 87.33 89.41	LR 90.97 88.51 87.76 82.61	LF 91.76 89.63 87.54 85.88	LM 37.83 26.04 17.21	LP 80.14 76.81 72.42 70.35	LR 75.79 70.72 72.37 67.33	LF 77.90 73.64 72.40 68.80	LM 10.68 07.12 06.82 05.42



New in 2016: CCG Word–Word Dependencies

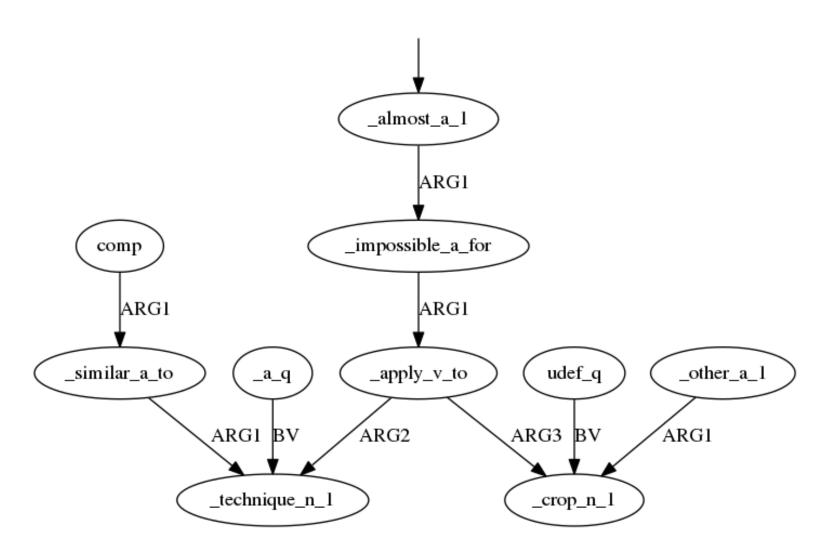


CCD: Canonical Conversion from CCGbank

- Connect lexical dependencies with properties from derivation in CCGbank;
- CCG categories as 'frame' identifiers; edge labels identify argument position.



Closer to Home: Abstract Semantic Graphs (E.g. EDS)





Biological Event Extraction (Björne, et al., 2009)

•



Biological Event Extraction (Björne, et al., 2009)

Negation Scope and Focus (Lapponi, et al., 2012)



Biological Event Extraction (Björne, et al., 2009)

•

Negation Scope and Focus (Lapponi, et al., 2012)

lacktriangle

Fine-Grained Opinion Analysis (Johansson & Moschitti, 2013)

•



Biological Event Extraction (Björne, et al., 2009)

•

Negation Scope and Focus (Lapponi, et al., 2012)

lacktriangle

Fine-Grained Opinion Analysis (Johansson & Moschitti, 2013)

Initial Set: Three (Nearly) SotA Systems Assumed to Benefit from Parsing.



```
But {this theory would} \ \( not \) {work}.

I think, Watson, {a brandy and soda would do him} \ \( no \) {harm}.

They were all confederates in {the same} \ \( un \) {known crime}.

"Found dead \( without \) {a mark upon him}.
```



```
But {this theory would} ⟨not⟩ {work}.

I think, Watson, {a brandy and soda would do him} ⟨no⟩ {harm}.

They were all confederates in {the same} ⟨un⟩ {known crime}.

"Found dead ⟨without⟩ {a mark upon him}.

{We have} ⟨never⟩ {gone out ⟨without⟩ {keeping a sharp watch}},

and ⟨no⟩ {one could have escaped our notice}."
```



```
But {this theory would} ⟨not⟩ {work}.

I think, Watson, {a brandy and soda would do him} ⟨no⟩ {harm}.

They were all confederates in {the same} ⟨un⟩ {known crime}.

"Found dead ⟨without⟩ {a mark upon him}.

{We have} ⟨never⟩ {gone out ⟨without⟩ {keeping a sharp watch}},

and ⟨no⟩ {one could have escaped our notice}."
```

Morante et al. (2011); Morante & Daelemans (2012)

- Fresh annotation of negation cues and their (possibly discontinous) scopes;
- semantics: "Scope of negation is the part of the meaning that is negated [...]"



```
But {this theory would} ⟨not⟩ {work}.

I think, Watson, {a brandy and soda would do him} ⟨no⟩ {harm}.

They were all confederates in {the same} ⟨un⟩ {known crime}.

"Found dead ⟨without⟩ {a mark upon him}.

{We have} ⟨never⟩ {gone out ⟨without⟩ {keeping a sharp watch}},

and ⟨no⟩ {one could have escaped our notice}."
```

Morante et al. (2011); Morante & Daelemans (2012)

- Fresh annotation of negation cues and their (possibly discontinous) scopes;
- semantics: "Scope of negation is the part of the meaning that is negated [...]"

Phorbol activation was positively modulated by Ca2+ influx while {TNF alpha activation was} (not).



Interchange Format for Syntactico-Semantic Graphs



Participating Teams and Approaches



Preliminary Results: Many Dimensions of Variation



Very Much in the Making these Days ...

http://epe.nlpl.eu

