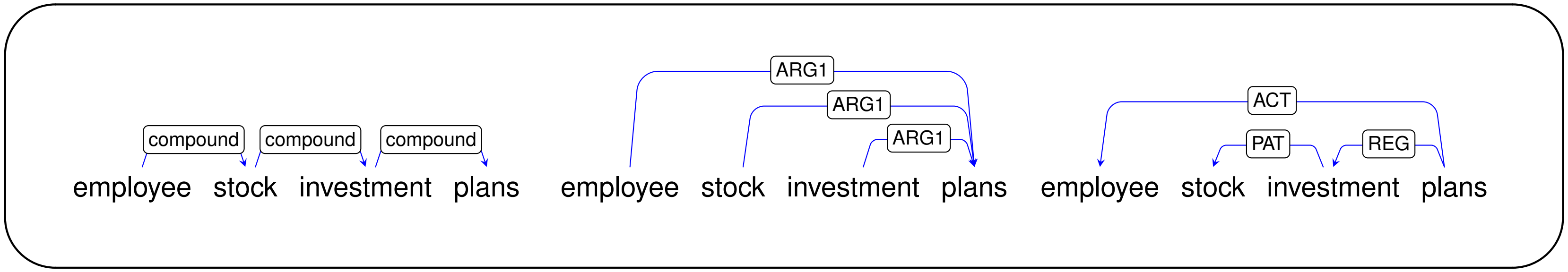


EPE 2017

Towards an Infrastructure for Extrinsic Parser Evaluation

Stephan Oepen

Universitetet i Oslo and
Center for Advanced Study at the Norwegian Academy of Science and Letters

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

Two Decades 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: 84.2 (Magerman, 1995) $\rightarrow$ 91.0 (Charniak & Johnson, 2005);
- some ten years later, neural advances: 93.8 (Choe & Charniak, 2016).

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, maybe, 1960s)

Two Decades 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: 84.2 (Magerman, 1995) $\rightarrow$ 91.0 (Charniak & Johnson, 2005);
- some ten years later, neural advances: 93.8 (Choe & Charniak, 2016).

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, maybe, 1960s)

Two Decades 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: 84.2 (Magerman, 1995) $\rightarrow$ 91.0 (Charniak & Johnson, 2005);
- some ten years later, neural advances: 93.8 (Choe & Charniak, 2016).

Parallel Contributions to Natural Language 'Understanding'?

Extrinsic Evaluation: Motivation & Goals

Limitations in Intrinsic Evaluation

- Presupposes ‘gold-standard’ syntactico-semantic target representations;
- out of necessity, typically limited to narrow range of domains and genres;
- repeated testing (sometimes over decades) against the same benchmark;
- granular output similarity metrics (e.g. ParsEval or LAS) hard to interpret;
- and maybe mis-leading: one mis-attachment can make all the difference.

Extrinsic Evaluation: Motivation & Goals

Limitations in Intrinsic Evaluation

- Presupposes 'gold-standard' syntactico-semantic target representations;
- out of necessity, typically limited to narrow range of domains and genres;
- repeated testing (sometimes over decades) against the same benchmark;
- granular output similarity metrics (e.g. ParsEval or LAS) hard to interpret;
- and maybe mis-leading: one mis-attachment can make all the difference.

Desiderata for Extrinsic Parser Evaluation

- Informative about *downstream utility* for broad range of NLU applications;
- applicable across *diverse* output representations and parsing approaches;
- easy to reproduce and apply with new parsers, for *all parser developers*.

The EPE 2017 Shared Task: What We Did

The EPE 2017 Shared Task: What We Did

(0) Team up with developers of relevant downstream systems;

The EPE 2017 Shared Task: What We Did

(0) Team up with developers of relevant downstream systems;

(1) Select (publicly available) data sets and evaluation metrics;

The EPE 2017 Shared Task: What We Did

(0) Team up with developers of relevant downstream systems;

(1) Select (publicly available) data sets and evaluation metrics;

(2) Define generalized notion of ‘dependency representations’;

The EPE 2017 Shared Task: What We Did

(0) Team up with developers of relevant downstream systems;

(1) Select (publicly available) data sets and evaluation metrics;

(2) Define generalized notion of ‘dependency representations’;

(3) Uniform interchange format as common parser interface;

The EPE 2017 Shared Task: What We Did

(0) Team up with developers of relevant downstream systems;

(1) Select (publicly available) data sets and evaluation metrics;

(2) Define generalized notion of ‘dependency representations’;

(3) Uniform interchange format as common parser interface;

(4) Make three state-of-the-art systems robust to divergence;

The EPE 2017 Shared Task: What We Did

(0) Team up with developers of relevant downstream systems;

(1) Select (publicly available) data sets and evaluation metrics;

(2) Define generalized notion of ‘dependency representations’;

(3) Uniform interchange format as common parser interface;

(4) Make three state-of-the-art systems robust to divergence;

(5) Automated re-training for each submitted parser output;

The EPE 2017 Shared Task: What We Did

(0) Team up with developers of relevant downstream systems;

(1) Select (publicly available) data sets and evaluation metrics;

(2) Define generalized notion of ‘dependency representations’;

(3) Uniform interchange format as common parser interface;

(4) Make three state-of-the-art systems robust to divergence;

(5) Automated re-training for each submitted parser output;

(6) Low barrier to participation: Run your parser on our text.

Extrinsic Evaluation: Methodological Challenges

Tease Apart Various Contributions

- Parser is one component in complex end-to-end systems; does it matter?

→ pick applications ‘sensitive’ to grammatical structure: hierarchical events;

→ contrast state-of-the-art parser outputs with ‘baseline’ dependency graphs.

Extrinsic Evaluation: Methodological Challenges

Tease Apart Various Contributions

- Parser is one component in complex end-to-end systems; does it matter?

→ pick applications ‘sensitive’ to grammatical structure: hierarchical events;

→ contrast state-of-the-art parser outputs with ‘baseline’ dependency graphs.

Informative & Plausible Measurements

- Evaluate at state-of-the-art performance levels (even if a moving target);

→ EPE 2017 end-to-end performances more than competitive with prior art.

Extrinsic Evaluation: Methodological Challenges

Tease Apart Various Contributions

- Parser is one component in complex end-to-end systems; does it matter?

$\rightarrow$ pick applications 'sensitive' to grammatical structure: hierarchical events;

$\rightarrow$ contrast state-of-the-art parser outputs with 'baseline' dependency graphs.

Informative & Plausible Measurements

- Evaluate at state-of-the-art performance levels (even if a moving target);

$\rightarrow$ EPE 2017 end-to-end performances more than competitive with prior art.

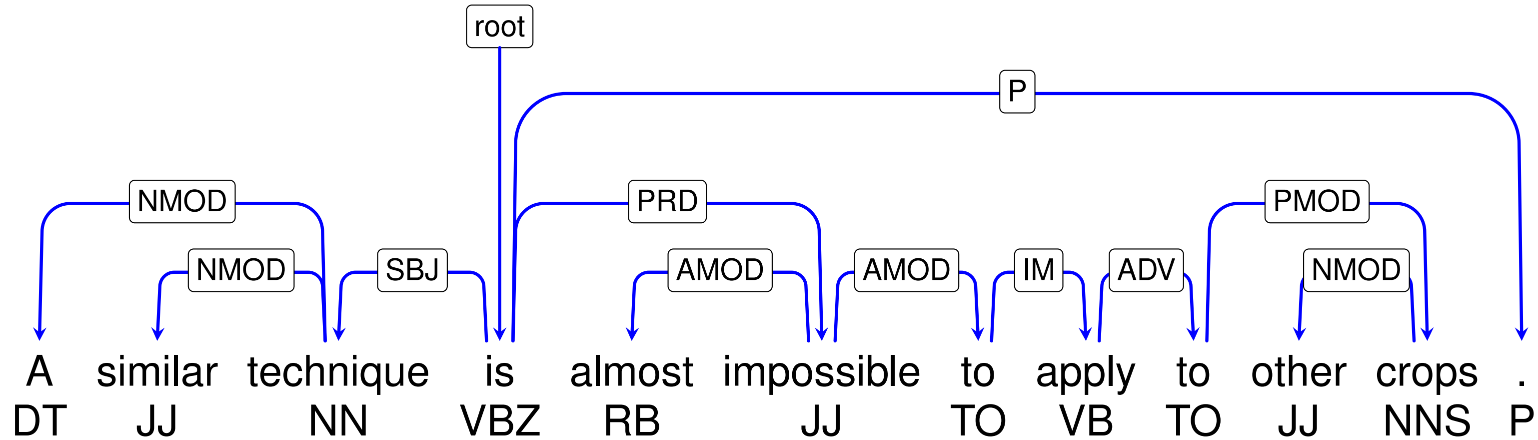
No 'Bias' Towards Individual Analysis Schemes

$\rightarrow$ Automatic re-training of downstream systems; input 'pseudonymizaton';

$\rightarrow$ feature engineering and tuning originally only against one type of inputs.

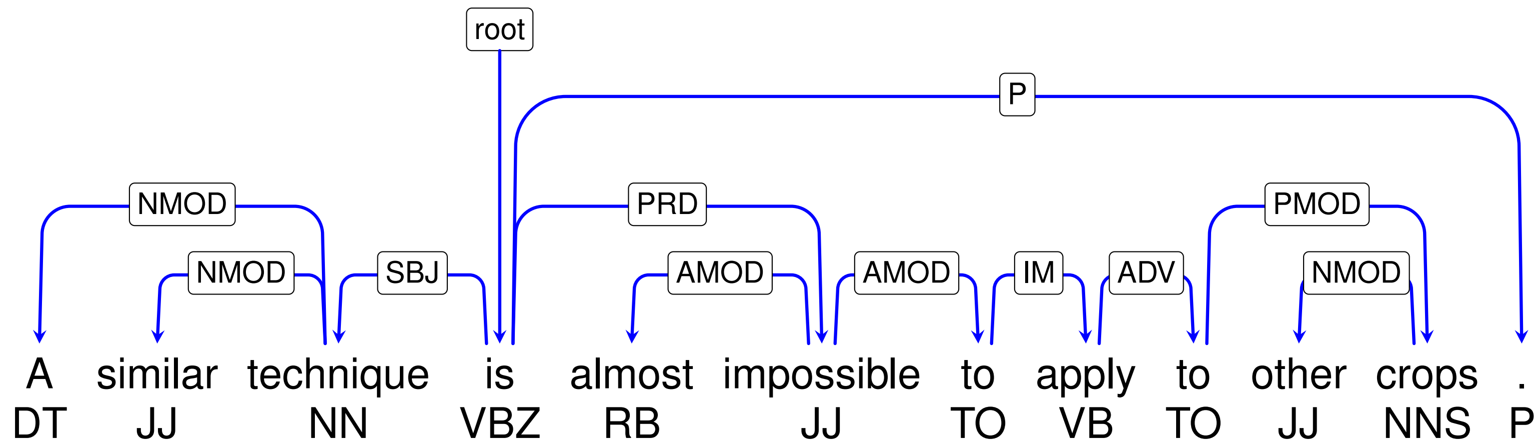
A Sample of Syntactic Dependencies

CoNLL

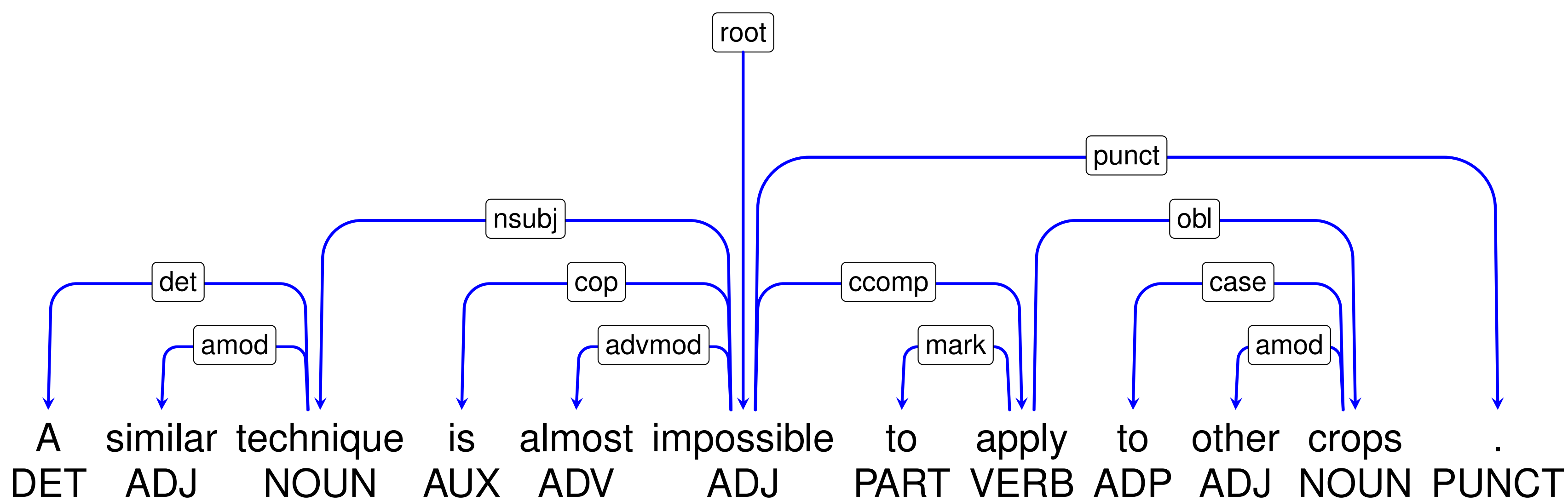


A Sample of Syntactic Dependencies

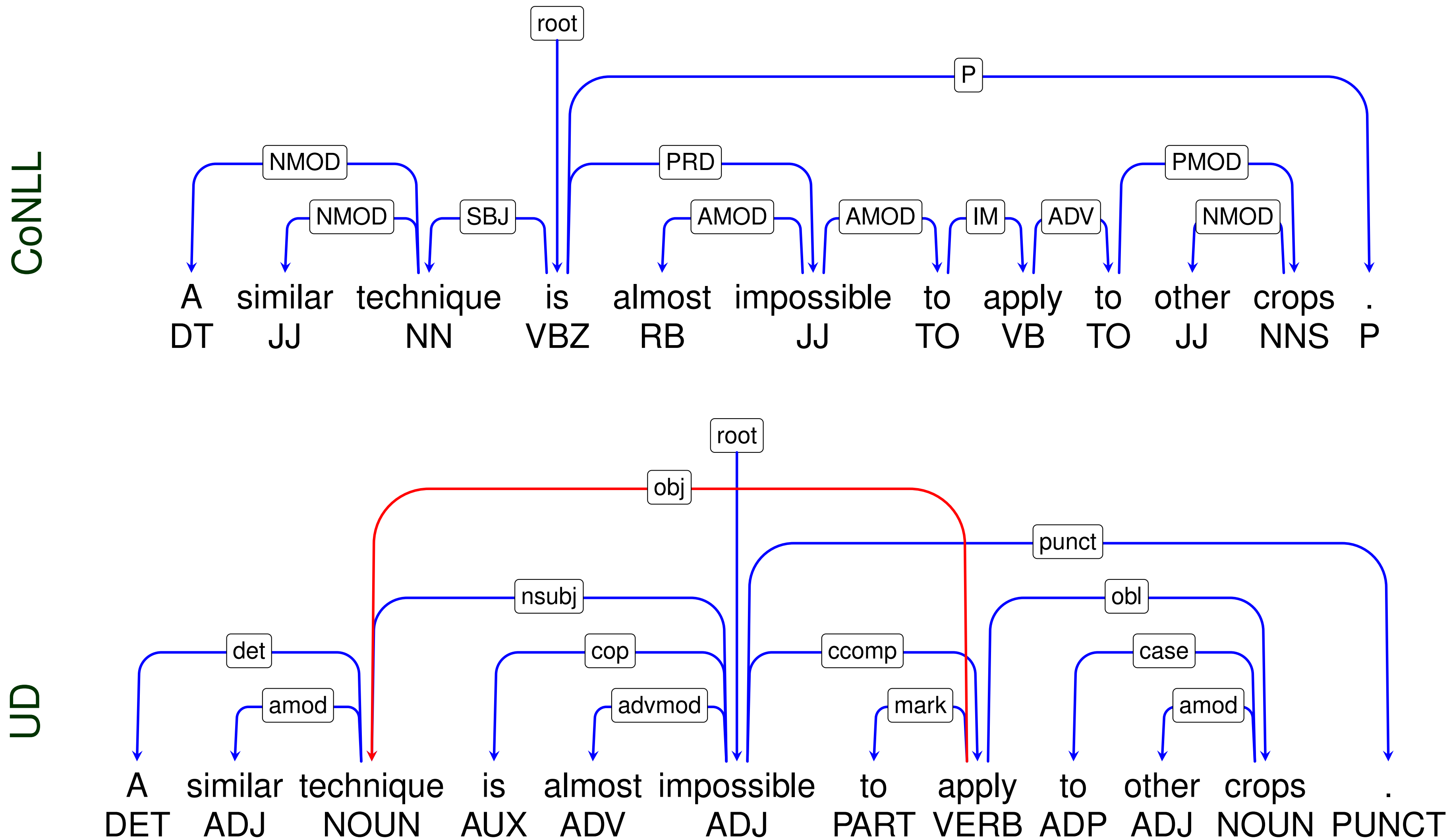
CoNLL



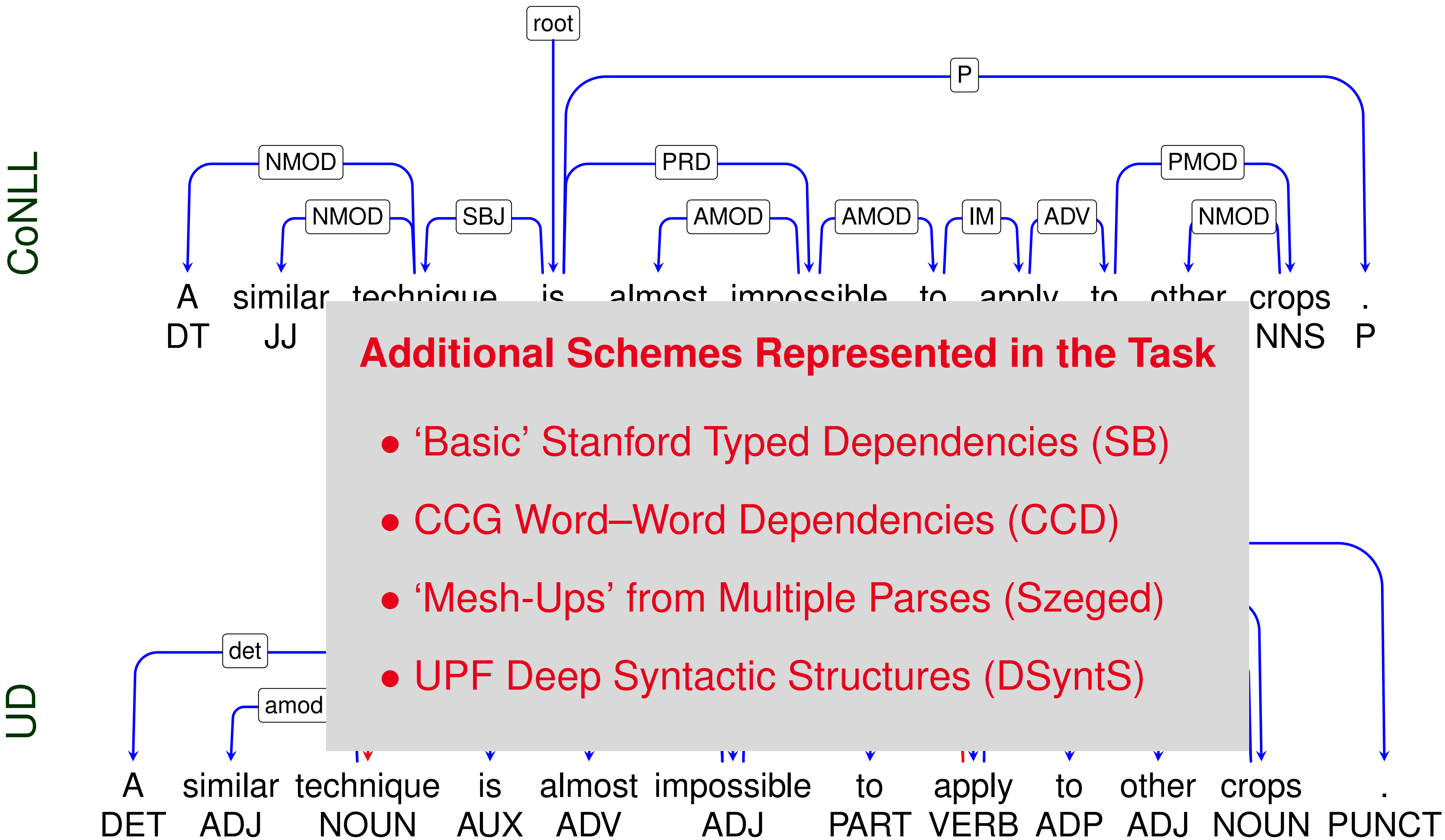
UD



A Sample of Syntactic Dependencies



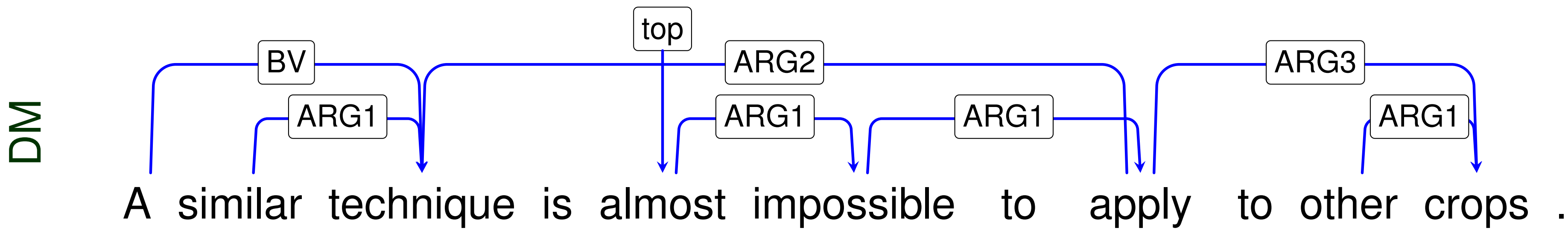
A Sample of Syntactic Dependencies



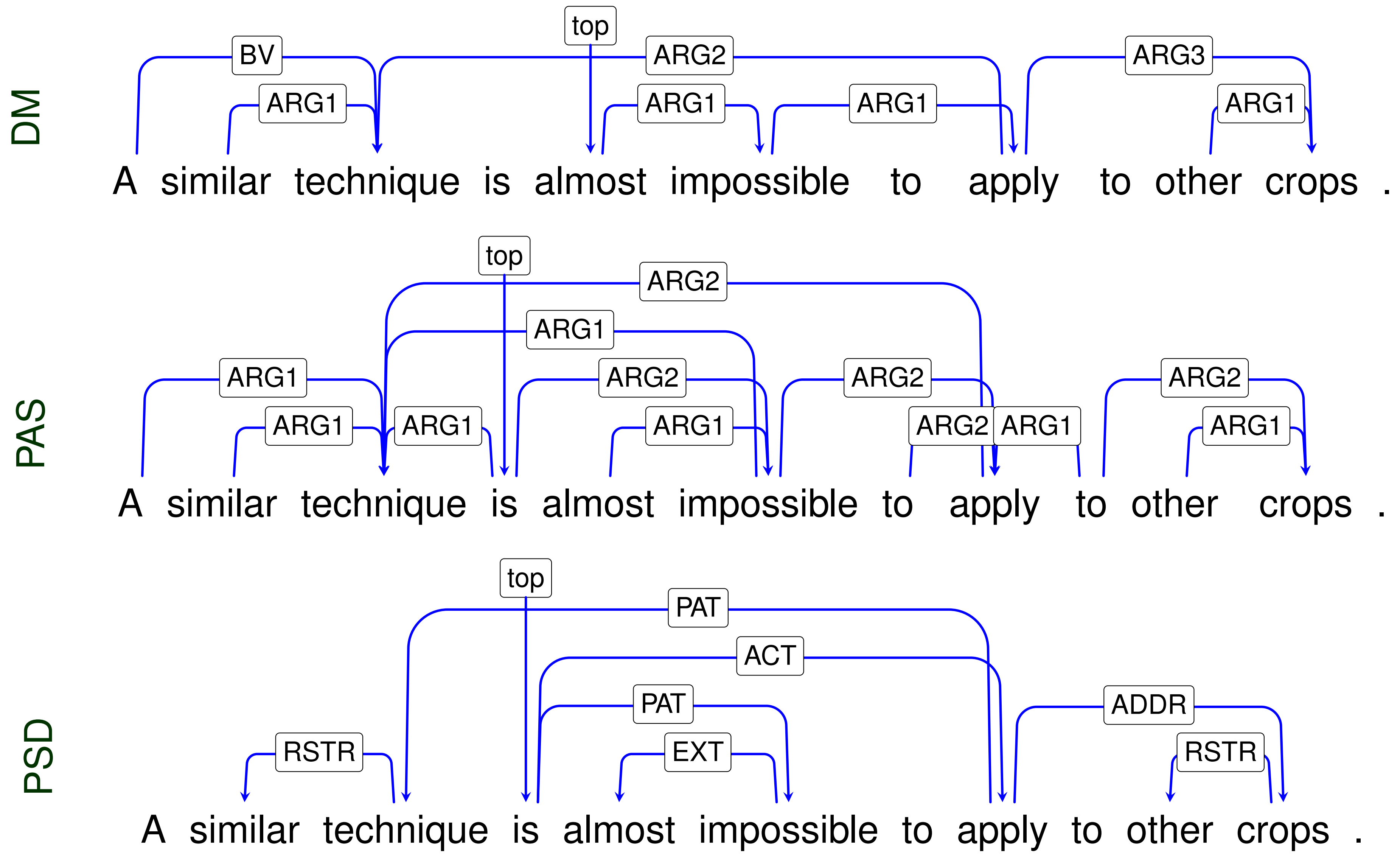
Additional Schemes Represented in the Task

- ‘Basic’ Stanford Typed Dependencies (SB)
- CCG Word–Word Dependencies (CCD)
- ‘Mesh-Ups’ from Multiple Parses (Szeged)
- UPF Deep Syntactic Structures (DSyntS)

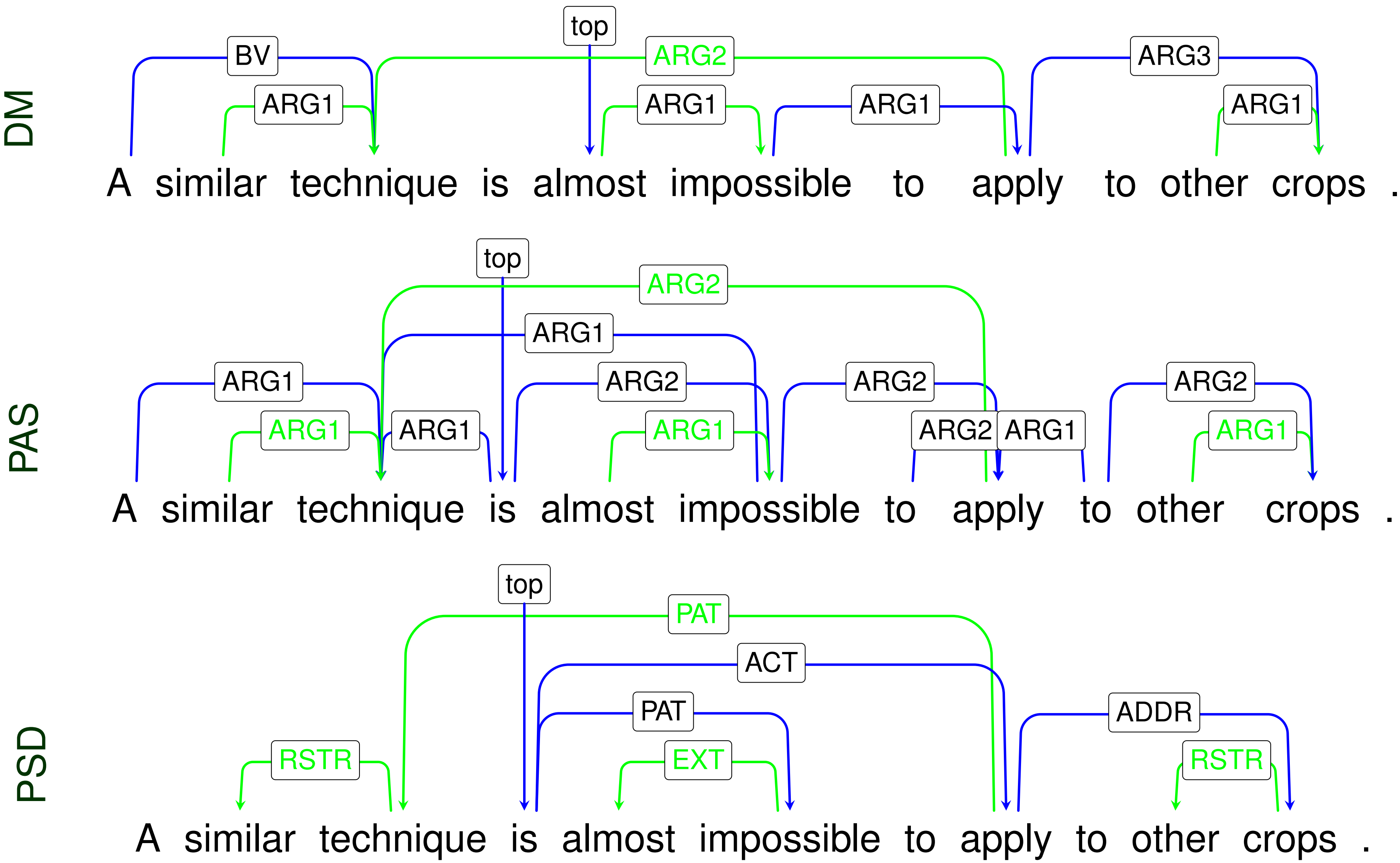
SDP: Bi-Lexical Semantic Dependencies



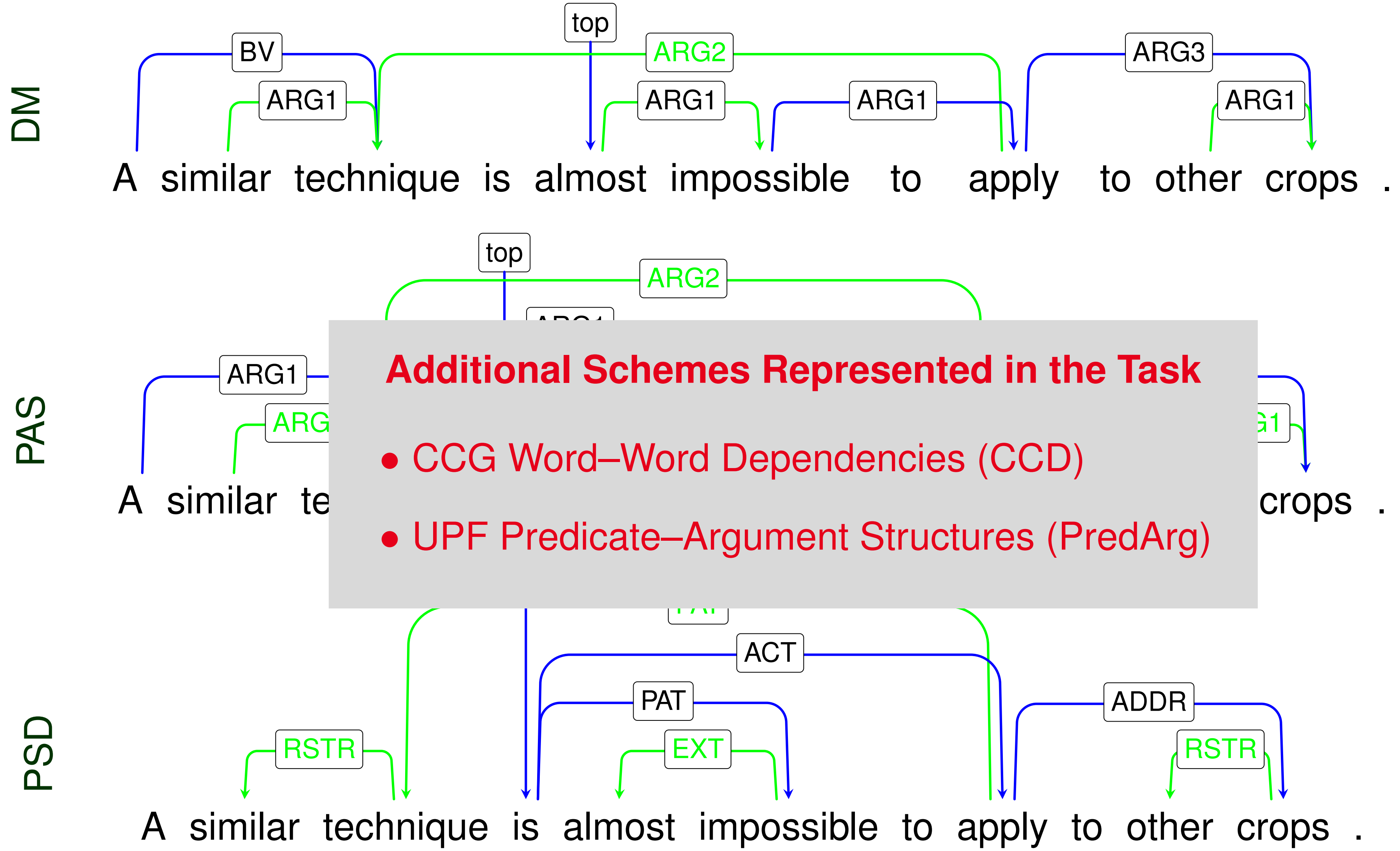
SDP: Bi-Lexical Semantic Dependencies



SDP: Bi-Lexical Semantic Dependencies



SDP: Bi-Lexical Semantic Dependencies



Additional Schemes Represented in the Task

- CCG Word–Word Dependencies (CCD)
- UPF Predicate–Argument Structures (PredArg)

Major Dimensions of Variation

EPE 2017 Limits Itself to English Dependency Parsing

Major Dimensions of Variation

EPE 2017 Limits Itself to English Dependency Parsing

Formal Graph Properties

- Rooted trees vs. general directed graphs: node re-entrancies; singletons;
- unique *root* node with zero in-degree vs. zero to n (semantic) *top* nodes.

Major Dimensions of Variation

EPE 2017 Limits Itself to English Dependency Parsing

Formal Graph Properties

- Rooted trees vs. general directed graphs: node re-entrancies; singletons;
- unique *root* node with zero in-degree vs. zero to n (semantic) *top* nodes.

Linguistic Design Decisions

- Function vs. 'content' words as heads: e.g. auxiliaries and prepositions;
- directionality: e.g. determiners and adjectives as predicates semantically.

Major Dimensions of Variation

EPE 2017 Limits Itself to English Dependency Parsing

Formal Graph Properties

- Rooted trees vs. general directed graphs: node re-entrancies; singletons;
- unique *root* node with zero in-degree vs. zero to n (semantic) *top* nodes.

Linguistic Design Decisions

- Function vs. 'content' words as heads: e.g. auxiliaries and prepositions;
- directionality: e.g. determiners and adjectives as predicates semantically.

Pushing the Notion of Lexicalization

- Relax one-to-one correspondence to tokens: 'empty' or overlapping nodes.

Interchange Format for Syntactico-Semantic Graphs

The term (bi-lexical) dependency representation in the context of EPE 2017 is interpreted as a graph whose nodes are anchored in surface lexical units, and whose edges represent labeled directed relations between two nodes. Each node corresponds to a sub-string of the underlying linguistic signal (input string), identified by character stand-off pointers. Node labels can comprise a non-recursive attribute–value matrix (or 'feature structure'), for example to encode lemma and part of speech information. Each graph can optionally designate one or more 'top' nodes, broadly interpreted as the root-level head or highest-scoping predicate. [Oepen et al., 2017]

Interchange Format for Syntactico-Semantic Graphs

The term (bi-lexical) dependency representation in the context of EPE 2017 is interpreted as a graph whose nodes are anchored in surface lexical units, and whose edges represent labeled directed relations between two nodes. Each node corresponds to a sub-string of the underlying linguistic signal (input string), identified by character stand-off pointers. Node labels can comprise a non-recursive attribute–value matrix (or 'feature structure'), for example to encode lemma and part of speech information. Each graph can optionally designate one or more 'top' nodes, broadly interpreted as the root-level head or highest-scoping predicate. [Oepen et al., 2017]

- Allow divergent segmentations: stand-off annotations; not token-centric;
- graph serialization in JSON: human- & machine-readable; easy to extend.

EPE 2017: Supported Downstream Applications

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

- Hierarchically nested event triggers, each with its arguments and modifiers.

EPE 2017: Supported Downstream Applications

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

- Hierarchically nested event triggers, each with its arguments and modifiers.

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

- Negation cues, with partly overlapping, discontinuous scopes and focus.

EPE 2017: Supported Downstream Applications

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

- Hierarchically nested event triggers, each with its arguments and modifiers.

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

- Negation cues, with partly overlapping, discontinuous scopes and focus.

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

- Partly overlapping opinion expressions, each with opinion holder and polarity.

EPE 2017: Supported Downstream Applications

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

- Hierarchically nested event triggers, each with its arguments and modifiers.

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

- Negation cues, with partly overlapping, discontinuous scopes and focus.

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

- Partly overlapping opinion expressions, each with opinion holder and polarity.

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

Participating Teams and Approaches (1/2)

East China Normal University [5 Runs]

- Neural, transition-based parser (Kiperwasser & Goldberg, 2016)—UD.

Participating Teams and Approaches (1/2)

East China Normal University [5 Runs]

- Neural, transition-based parser (Kiperwasser & Goldberg, 2016)—UD.

INRIA, Paris Diderot, Paris Sorbonne (With Stanford) [12 Runs]

- Neural, transition-based tree-to-graph parser; two sets of training data;
- systematic variation of representations: SDP & many UD 'enhancements'.

Participating Teams and Approaches (1/2)

East China Normal University [5 Runs]

- Neural, transition-based parser (Kiperwasser & Goldberg, 2016)—UD.

INRIA, Paris Diderot, Paris Sorbonne (With Stanford) [12 Runs]

- Neural, transition-based tree-to-graph parser; two sets of training data;
- systematic variation of representations: SDP & many UD ‘enhancements’.

Peking University [6 Runs]

- Three different string-to-graph parsers; one of them neural—DM & CCD.

Participating Teams and Approaches (1/2)

East China Normal University [5 Runs]

- Neural, transition-based parser (Kiperwasser & Goldberg, 2016)—UD.

INRIA, Paris Diderot, Paris Sorbonne (With Stanford) [12 Runs]

- Neural, transition-based tree-to-graph parser; two sets of training data;
- systematic variation of representations: SDP & many UD ‘enhancements’.

Peking University [6 Runs]

- Three different string-to-graph parsers; one of them neural—DM & CCD.

Charles University in Prague [5 Runs]

- Variants of UDPipe system: representations; version; pre-processing—UD.

Participating Teams and Approaches (2/2)

University of Szeged [5 Runs]

- Integrating dependencies from multiple parsers and parses—CoNLL^{++}.

Participating Teams and Approaches (2/2)

University of Szeged [5 Runs]

- Integrating dependencies from multiple parsers and parses—CoNLL^{++}.

Universitat Pompeu Fabre [3 Runs]

- Three strata: surface syntax, ‘deep’ syntax, predicate–argument structure;
- ‘classic’ string-to-tree parser; hand-crafted graph transduction grammars.

Participating Teams and Approaches (2/2)

University of Szeged [5 Runs]

- Integrating dependencies from multiple parsers and parses—CoNLL^{++}.

Universitat Pompeu Fabre [3 Runs]

- Three strata: surface syntax, 'deep' syntax, predicate–argument structure;
- 'classic' string-to-tree parser; hand-crafted graph transduction grammars.

Stanford University (With Paris) [11 Runs]

- Neural string-to-tree parser; heuristic rules to 'enhance' and 'normalize'.

Participating Teams and Approaches (2/2)

University of Szeged [5 Runs]

- Integrating dependencies from multiple parsers and parses—CoNLL^{++}.

Universitat Pompeu Fabre [3 Runs]

- Three strata: surface syntax, ‘deep’ syntax, predicate–argument structure;
- ‘classic’ string-to-tree parser; hand-crafted graph transduction grammars.

Stanford University (With Paris) [11 Runs]

- Neural string-to-tree parser; heuristic rules to ‘enhance’ and ‘normalize’.

University of Washington [1 Run]

- Neural, multi-task string-to-graph parser (Peng et al., 2017)—DM (SotA).

EPE 2017 Mechanics: Facts and Figures

Schedule

Mid-March Release training and development data; EPE interchange format;

Throughout April Data updates; pre-processed variant and fornmat converter;

Mid-April Pre-evaluation trial run: Five teams submitted 14 different runs;

Throughout June Debugging, with some teams; a couple of re-submissions;

Late July Final evaluation results; application and system descriptions;

September 20 Presentation of infrastructure, partcipants, and results;

...

An Ocean of Experimental Results

results.ods – LibreOffice Calc

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	team	run	representation	training	tokens	input	reference	Event Extraction			Negation Resolution			Opinion Analysis			Average	Rank
2								P	R	F1	P	R	F1	P	R	F1		
3	ecnu	0	UD v2.0	English 2.0	204,585	tt		49.48	39.00	43.62	99.17	45.45	62.33	60.27	57.42	58.81	**54.92**	
4	ecnu	1	UD v2.0	English 2.0	204,585	tt		50.72	38.97	44.08	99.17	45.45	62.33	62.86	60.04	61.42	**55.94**	
5	ecnu	2	UD v2.0	English 2.0	204,585	tt		52.24	40.23	45.46	99.17	45.45	62.33	62.15	59.75	60.93	**56.24**	**5**
6	ecnu	3	UD v2.0	English 2.0	204,585	tt		54.53	35.58	43.06	99.18	45.83	62.69	62.11	58.17	60.08	**55.28**	
7	ecnu	4	UD v2.0	English 2.0	204,585	tt		60.69	35.76	45.00	99.15	43.94	60.89	63.32	61.07	62.17	**56.02**	
8	oxford	0	EDS					0.00	0.00									
9	paris-stanford	0	DM	WSJ 00–20 (SDP Sub-Set)	802,717	txt		59.11	37.71	46.04	99.12	42.80	59.78	65.04	51.32	57.37	**54.40**	
10	paris-stanford	1	PAS	WSJ 00–20 (SDP Sub-Set)	802,717	txt		52.39	40.98	45.99	99.09	41.29	58.29	65.80	52.73	58.54	**54.27**	
11	paris-stanford	2	UD v1 basic	WSJ 00–20 (SDP Sub-Set)	802,717	txt		55.79	44.56	49.55	99.04	39.02	55.98	65.87	61.30	63.50	**56.34**	
12	paris-stanford	3	UD v1 enhanced	WSJ 00–20 (SDP Sub-Set)	802,717	txt		57.48	41.64	48.29	99.06	39.77	56.75	66.22	62.43	64.27	**56.44**	
13	paris-stanford	4	UD v1 enhanced++	WSJ 00–20 (SDP Sub-Set)	802,717	txt		58.55	39.50	47.17	99.03	38.64	55.59	65.10	61.75	63.38	**55.38**	
14	paris-stanford	5	UD v1 enhanced++ diathesis	WSJ 00–20 (SDP Sub-Set)	802,717	txt		55.58	43.37	48.72	99.03	38.64	55.59	66.62	62.03	64.24	**56.18**	
15	paris-stanford	6	UD v1 enhanced++ diathesis--	WSJ 00–20 (SDP Sub-Set)	802,717	txt		58.11	39.19	46.81	99.06	39.77	56.75	64.21	60.27	62.18	**55.25**	
16	paris-stanford	7	UD v1 basic	WSJ, Brown, GENIA	1,692,030	txt		57.69	42.80	49.14	99.05	39.39	56.36	65.78	60.96	63.28	**56.26**	
17	paris-stanford	8	UD v1 enhanced	WSJ, Brown, GENIA	1,692,030	txt		54.90	44.75	49.31	99.07	40.15	57.14	65.59	62.42	63.97	**56.81**	**3**
18	paris-stanford	9	UD v1 enhanced++	WSJ, Brown, GENIA	1,692,030	txt		58.03	43.02	49.41	99.04	39.02	55.98	66.77	61.04	63.78	**56.39**	
19	paris-stanford	10	UD v1 enhanced++ diathesis	WSJ, Brown, GENIA	1,692,030	txt		59.88	40.19	48.10	98.97	36.36	53.18	65.86	60.92	63.29	**54.86**	
20	paris-stanford	11	UD v1 enhanced++ diathesis--	WSJ, Brown, GENIA	1,692,030	txt		58.92	40.07	47.70	99.06	39.77	56.75	64.90	60.56	62.65	**55.70**	
21	peking	0	DM	SDP 2015	802,717	tt		59.28	34.22	43.39	99.15	43.94	60.89	65.63	53.64	59.03	**54.44**	
22	peking	1	CCD	SDP 2016	801,149	tt		58.26	40.07	47.48	99.15	44.32	61.26	66.57	54.55	59.96	**56.23**	**6**
23	peking	2	DM	SDP 2015	802,717	tt												
24	peking	3	CCD	SDP 2016	801,149	tt												
25	peking	4	DM	SDP 2015	802,717	tt		55.42	40.95	47.10	99.10	41.67	58.67	65.74	53.66	59.09	**54.95**	
26	peking	5	CCD	SDP 2016	801,149	tt		54.73	42.17	47.64	99.12	42.42	59.41	66.97	54.84	60.30	**55.78**	
27	prague	0	UD v2.0	English 2.0	204,585	txt	CoNLL 2017	53.84	36.61	43.58	99.10	41.83	58.83	62.61	57.21	59.79	**54.07**	
28	prague	1	UD v2.0	English 2.0	204,585	tt		56.35	38.21	45.54	99.16	44.70	61.62	62.31	59.74	61.00	**56.05**	**7**
29	prague	2	UD v2.0	English, LinES, ParTUT 2.0	292,205	txt		53.22	37.87	44.25	99.12	42.97	59.95	63.45	54.63	58.71	**54.30**	
30	prague	3	UD v2.0	English 2.0	192,552	txt	CoNLL 2017	51.91	36.27	42.70	99.12	42.97	59.95	61.26	56.72	58.90	**53.85**	
31	prague	4	UD v1.2	English 2.0	204,585	txt		51.71	37.12	43.22	98.90	34.22	50.85	61.00	56.25	58.53	**50.86**	
32	stanford-paris	0	Stanford Basic	WSJ, Brown, GENIA	1,692,030	txt		56.93	45.03	50.29	99.22	48.48	65.13	67.26	60.54	63.72	**59.71**	
33	stanford-paris	1	UD v1 basic	WSJ 00–20 (SDP Sub-Set)	802,717	txt		57.59	40.76	47.73	99.19	46.21	63.05	67.47	61.30	64.24	**58.34**	
34	stanford-paris	2	UD v1 enhanced	WSJ 00–20 (SDP Sub-Set)	802,717	txt		57.24	40.98	47.76	99.20	46.97	63.75	67.69	61.02	64.18	**58.57**	
35	stanford-paris	3	UD v1 enhanced++	WSJ 00–20 (SDP Sub-Set)	802,717	txt		56.76	42.74	48.76	99.21	47.35	64.10	67.43	61.58	64.37	**59.08**	
36	stanford-paris	4	UD v1 enhanced++ diathesis	WSJ 00–20 (SDP Sub-Set)	802,717	txt		58.86	40.51	47.99	99.19	46.21	63.05	66.68	61.95	64.23	**58.42**	

Overall / Event Extraction / Negation Resolution / Opinion Analysis

All Available: Data, Systems, Results, Scores

http://epe.nlpl.eu

All Available: Data, Systems, Results, Scores

`http://epe.nlpl.eu`

Looking for collaborators: parser and application developers.