Mod-D2T: A Multi-layer Dataset for Modular Data-to-Text Generation

Simon Mille *simon.mille@adaptcentre.ie*
Fançois Lareau *francois.lareau@umontreal.ca*
Stamatia Dasiopoulou *stamatia.dasiopoulou@gmail.com*
Anya Belz *anya.belz@adaptcentre.ie*

Why this dataset?

1. Splitting generation in substeps recently gave good results.
2. There is no rich modular dataset for D2T.

What for

- Learning NLG/NLU modules with any pair of levels.
- Make possible combinations of modules of different implementations
- Teaching, etc.

How was this dataset built?

We run the FORGe rule-based generator on the WebNLG dataset and cleaned all intermediate layers.

Synthetic dataset: Pros/cons

+ Possibility to generate many versions of the data
+ Parallel data in several languages
+ Quality of intermediate representations
- Variety and quality of texts is limited

The layers of our dataset

Reiter&Dale Tasks	Mod-D2T Tasks	Mod-D2T Input	Mod-D2T Output	Structure type
Content determination	—	—	—	
Discourse planning	Linguistic structuring	WebNLG	PredArg	Directed acyclic graphs
Sentence aggregation	Text planning*	PredArg	PredArg-Agg	Directed acyclic graphs
	Lexicalisation	PredArg(-Agg)	PredArg-Lex	Directed acyclic graphs
	Comm. structuring	PredArg-Lex	PredArg-Th	Directed acyclic graphs
Lexicalisation	Deep sent. structuring	PredArg-Th	DSynt	Dependency trees
	Surf. sent. structuring	DSynt	SSynt	Dependency trees
	Synt. aggregation*	SSynt	SSynt-Agg	Dependency trees
REG	REG*	SSynt(-Agg)	SSynt-Pro	Dependency trees
Linguistic realisation	Word ord. and agree. resolution	SSynt(-Agg/-Pro)	DMorph	Chains
	Surface form retrieval	DMorph	SMorph	Chains

WebNLG data point

```
<entry category="Building" eid="Id10" shape="(x (x) (x) (x) (x))"
    shape_type="sibling" size="4">
  <modifiedtripleset>
    <mtriple> 103_Colmore_Row | location | Birmingham </mtriple>
    <mtriple> 103_Colmore_Row | architect | Jonh_Madin </mtriple>
    <mtriple> 103_Colmore_Row | floorCount | 23 </mtriple>
    <mtriple> 103_Colmore_Row | completionDate | 1976 </mtriple>
  </modifiedtripleset>
</entry>
```

Generated Text

103 Colmore Row, designed by John Madin, is in Birmingham. It has 23 floors and was completed in 1976.

10 intermediate representations in CoNLL-U format below: DAGs, DGs, Chains

PredArg

ID	Semanteme	Features	Head	Rel	Misc
1	be	_	0	root	src=1
2	Birmingham	location|ne	1	A2	src=3
3	103_Colmore_Row	ne	1	A1	src=2|coref=0
4	<SENT>	_	_	_	_
5	design	past	0	root	src=4
6	John_Madin	person|ne	5	A1	src=6
7	103_Colmore_Row	ne	5	A2	src=5|coref=0
8	<SENT>	_	_	_	_
9	have	_	0	root	src=7
10	23	_	0	root	src=9
11	floor	_	9,10	A2,A1	src=10
12	103_Colmore_Row	ne	9	A1	src=8|coref=0
13	<SENT>	_	_	_	_
14	complete	past	0	root	src=12
15	1976	year|ne	14	Time	src=14
16	103_Colmore_Row	ne	14	A2	src=13|coref=0
17	<SENT>	_	_	_	_

PredArg-Agg

ID	Semanteme	Features	Head	Rel	Misc
1	be	_	0	root	src=1
2	design	past	0	root	src=4
3	103_Colmore_Row	ne	**1,2**	**A1,A2**	src=2|coref=0
4	Birmingham	location|ne	1	A2	src=3
5	John_Madin	person|ne	2	A1	src=6
6	<SENT>	_	_	_	_
7	23	_	0	root	src=9
8	have	_	0	root	src=7
9	floor	_	7,8	A1,A2	src=10
10	103_Colmore_Row	ne	8	A1	src=8|coref=0
11	<SENT>	_	_	_	_
12	complete	past	0	root	src=12
13	1976	year|ne	12	Time	src=14
14	103_Colmore_Row	ne	12	A2	src=13|coref=0
15	<SENT>	_	_	_	_

PredArg-Lex

ID	Lexeme	POS	Features	Head	Rel	Misc
1	be	**VB**	_	0	root	src=1
2	Birmingham	**NP**	location|ne	1	A2	src=3
3	design	**VB**	past	0	root	src=4
4	John_Madin	**NP**	person|ne	3	A1	src=6
5	103_Colmore_Row	**NP**	ne	1,3	A1,A2	src=2|coref=0
6	<SENT>	_	_	_	_	_
7	23	**CD**	_	0	root	src=9
8	have	**VB**	_	0	root	src=7
9	floor	**NN**	_	7,8	A1,A2	src=10
10	103_Colmore_Row	**NP**	ne	8	A1	src=8|coref=0
11	<SENT>	_	_	_	_	_
12	complete	**VB**	past	0	root	src=12
13	1976	**NP**	year|ne	12	Time	src=14
14	103_Colmore_Row	**NP**	ne	12	A2	src=13|coref=0
15	<SENT>	_	_	_	_	_

PredArg-Th

ID	Lexeme	POS	Features	Head	Rel	Misc
1	design	VB	past	0	root	src=4
2	John_Madin	NP	person|ne	1	A0	src=6
3	be	VB	**rheme**	0	root	src=1
4	103_Colmore_Row	NP	ne	3,1	A1,A1	src=2|coref=0
5	Birmingham	NP	location|ne	3	A2	src=3
6	<SENT>	_	_	_	_	_
7	23	CD	_	0	root	src=9
8	have	VB	**rheme**	0	root	src=7
9	floor	NN	_	7,8	A1,A2	src=10
10	103_Colmore_Row	NP	ne	8	A1	src=8|coref=0
11	<SENT>	_	_	_	_	_
12	point_time_year	_	_	0	root	src=14
13	1976	NP	year|ne	12	A2	src=14
14	complete	VB	past|**rheme**	12	A1	src=12
15	103_Colmore_Row	NP	ne	14	A1	src=13|coref=0
16	<SENT>	_	_	_	_	_

DSynt

ID	Lexeme	POS	Features	Head	Rel	Misc
1	be	VB	**decl|act|fin|pres**|rheme	0	root	src=1
2	Birmingham	NP	location	1	**II**	src=3
3	103_Colmore_Row	NP	_	**1**	**I**	src=2|coref=0
4	design	VB	**part**|past	**3**	**ATTR**	src=4
5	John_Madin	NP	person	4	**I**	src=6
6	<SENT>	_	_	_	_	_
7	have	VB	**decl|act|fin|pres**|rheme	0	root	src=7
8	floor	NN	**pl**	**7**	**II**	src=10
9	103_Colmore_Row	NP	_	7	**I**	src=8|coref=0
10	23	CD	_	**8**	**ATTR**	src=9
11	<SENT>	_	_	_	_	_
12	complete	VB	**decl|pass|fin**|past|rheme	0	root	src=12
13	in	IN	_	12	**ATTR**	src=14
14	1976	NP	year	13	**II**	src=14
15	103_Colmore_Row	NP	_	12	**II**	src=13|coref=0
16	<SENT>	_	_	_	_	_

SSynt

ID	Lexeme	POS	Features	Head	Rel	Misc
1	be	VB	decl|fin|**ind**|pres|rheme	0	root	src=1
2	103_Colmore_Row	NP	**sg**|ne	1	**SBJ**	src=2|coref=0
3	design	JJ	part	2	**NMOD**	src=4
4	in	IN	_	1	**PRD**	src=3
5	Birmingham	NP	**sg**|location|ne	4	**PMOD**	src=3
6	by	IN	_	3	**LGS**	src=6
7	John_Madin	NP	**masc|sg**|person|ne	6	**PMOD**	src=6
8	<SENT>	_	_	_	_	_
9	have	VB	decl|fin|**ind**|pres|rheme	0	root	src=7
10	floor	NN	**pl**	9	**OBJ**	src=10
11	23	CD		10	**NMOD**	src=9
12	103_Colmore_Row	NP	**sg**|ne	9	**SBJ**	src=8|coref=0
13	<SENT>	_	_	_	_	_
14	be	VB	decl|fin|**ind**|past|rheme	0	root	src=12
15	in	IN	_	14	**ADV**	src=14
16	1976	NP	year|ne	15	**PMOD**	src=14
17	103_Colmore_Row	NP	**sg**|ne	14	**SBJ**	src=13|coref=0
18	complete	VB	decl|**part**|rheme	14	**VC**	src=12
19	<SENT>	_	_	_	_	_

SSynt-Agg

ID	Lexeme	POS	Features	Head	Rel	Misc
1	be	VB	decl|fin|ind|pres|rheme	0	root	src=1
2	103_Colmore_Row	NP	sg|ne	1	SBJ	src=2|coref=0
3	design	JJ	part	2	NMOD	src=4
4	by	IN	_	3	LGS	src=6
5	in	IN	_	1	PRD	src=3
6	Birmingham	NP	sg|location|ne	5	PMOD	src=3
7	John_Madin	NP	masc|sg|person|ne	4	PMOD	src=6
8	<SENT>	_	_	_	_	_
9	have	VB	decl|fin|ind|pres|rheme	0	root	src=7
10	**and**	**CC**	_	**9**	**COORD**	src=-
11	be	VB	decl|fin|ind|past|rheme	10	CONJ	src=12
12	in	IN	_	11	ADV	src=14
13	1976	NP	year|ne	12	PMOD	src=14
14	103_Colmore_Row	NP	sg|ne	9	SBJ	src=8|coref=0
15	floor	NN	pl	9	OBJ	src=10
16	23	CD	_	15	NMOD	src=9
17	complete	VB	decl|part|rheme	11	VC	src=12
18	103_Colmore_Row	NP	sg|ne	11	SBJ	src=13|coref=0
19	<SENT>	_	_	_	_	_

SSynt-Pro

ID	Lexeme	POS	Features	Head	Rel	Misc
1	be	VB	decl|fin|ind|pres|rheme	0	root	src=1
2	103_Colmore_Row	NP	sg|ne	1	SBJ	src=2|coref=0
3	design	JJ	part	2	NMOD	src=4
4	by	IN	_	3	LGS	src=6
5	John_Madin	NP	masc|sg|person|ne	4	PMOD	src=6
6	in	IN	_	1	PRD	src=3
7	Birmingham	NP	sg|location|ne	6	PMOD	src=3
8	<SENT>	_	_	_	_	_
9	have	VB	decl|fin|ind|pres|rheme	0	root	src=7
10	and	CC	_	9	COORD	src=-
11	be	VB	decl|fin|ind|past|rheme	10	CONJ	src=12
12	in	IN	_	11	ADV	src=14
13	1976	NP	year|ne	12	PMOD	src=14
14	floor	NN	pl	9	OBJ	src=10
15	23	CD	_	14	NMOD	src=9
16	**_PRO_**	**PP**	sg|ne	9	SBJ	src=8|coref=0
17	complete	VB	decl|part|rheme	11	VC	src=12
18	**_PRO_**	**PP**	sg|ne	11	SBJ	src=13|coref=0
19	<SENT>	_	_	_	_	_

DMorph

ID	Word	POS	Features	Misc
1	103_Colmore_Row	NP	_	src=2|coref=0
2	design	JJ	part	src=4
3	by	IN	_	src=6
4	John_Madin	NP	_	src=6
5	be	VB	decl|fin|ind|pres|**sg**	src=1
6	in	IN	_	src=3
7	Birmingham	NP	_	src=3
8	.	_	_	src=-
9	_PRO_	PP	sg	src=8|coref=0
10	have	VB	decl|fin|ind|pres|**sg**	src=7
11	23	CD	_	src=9
12	floor	NN	pl	src=10
13	and	CC	_	src=-
14	_PRO_	PP	sg|**delete**	src=13|coref=0
15	be	VB	decl|fin|ind|past|**sg**	src=12
16	complete	VB	decl|part	src=12
17	in	IN	_	src=14
18	1976	NP	_	src=14
19	.	_	_	src=-

SMorph

ID	Word	POS	Misc
1	103_Colmore_Row	NP	src=2|coref=0
2	,	_	src=-
3	**designed**	JJ	src=4
4	by	IN	src=6
5	John_Madin	NP	src=6
6	,	_	src=-
7	**is**	VB	src=1
8	in	IN	src=3
9	Birmingham	NP	src=3
10	.	_	src=-
11	**it**	PP	src=8|coref=0
12	**has**	VB	src=7
13	23	CD	src=9
14	**floors**	NN	src=10
15	and	CC	src=-
16	**was**	VB	src=12
17	**completed**	VB	src=12
18	in	IN	src=14
19	1976	NP	src=14
20	.	_	src=-

Total Nodes (N) and Sentences (S) per layer

Layer	N	S
PredArg	152,664	48,776
PredArg-Agg	134,188	31,204
PredArg-Lex	134,188	31,204
PredArg-Comm	143,448	31,204
DSynt	169,325	31,204
SSynt	219,962	31,204
SSynt-Agg	222,970	27,557
REG	220,218	27,557
DMorph	247,795	27,557
Text	268,267	27,557

Average Nodes (N) and Sentences (S) per layer

Layer	N	S	N/S
PredArg	9.2	2.9	3.1
PredArg-Agg	8.1	1.9	4.4
PredArg-Lex	8.1	1.9	4.4
PredArg-Th	8.6	1.9	4.7
DSynt	10.2	1.9	5.5
SSynt	13.2	1.9	7.1
SSynt-Agg	13.4	1.7	8.2
SSynt-Pro	13.2	1.7	8.1
DMorph	14.9	1.7	9.1
SMorph	16.1	1.7	9.9

Evaluation of the data quality

Manual assessment of the quality of:

1. **30 PredArg** data points (corresponding to 30 texts): 66.7% of data points without error
2. **30 SSynt-Pro** data points (corresponding to 30 texts) 93% of data point without error
3. **~180 Texts** (from WebNLG 2020)
   a. Data coverage 95.3 (human-written 95.5)
   b. Relevance 94.6 (94.1)
   c. Correctness 93.6 (93.4)
   d. Text structure: 87.0 (91.2)
   e. Fluency 82.7 (88.1).

Tagsets

PredArg

Label	Description	Example
A0–A6	*n*-th argument of a predicate or quasi-predicate	speak→ English
Location	location	born→ Paris
Time	time	build→ 1932
NonCore	inverted first argument of a predicate	runway→ second
Set	list of elements	and→ speak
Elaboration	(i) none of governor or dependent are argument of the other (ii) unknown argument slot	above me→ 610m

DSynt

Label	Description	Example
I–VI	*n*-th complement of a syntactic predicate	speak→ English
ATTR	modifier	runway→ second
COORD	coordination	staff members→ and
APPEND	parenthetical modifier	Hypermarcas Brazil→ (s.a.)

SSynt

Label	Description	Example
ADV	adverbial (broadly)	built→ in 1932
AMOD	argument or modifier of an adjective	similar→ to
AMOD_COMP	argument of a comparative adjective	higher→ than
COORD	between conjunct and conjunction	and→ speak
DEP	underspecified	—
EXT	prepositional object (not *to*)	ask→for
IM	infinitive marker	to→ ask
IOBJ	dative object (after OBJ)	give→ her
LGS	logical subject	owned→ by
NMOD	argument or modifier of a noun	runway→ fifth
OBJ	non-prepositional object	give→ medal
OPRD	prepositional object (*to*)	give→ to
PMOD	complement of a preposition	to→ her
PRD	predicative complement	be→ president
SBJ	syntactic subject	play→ Beatles
SUB	complement of a conjunction	while→ be

Future work

1. Create data for more languages: Irish, Spanish, French, etc.
2. Create variants of the dataset:
   a. with or without optional modules (* in table above).
   b. using different lexicalisations.
   c. using some ML modules to improve fluency (e.g. for Text Planning)

References

FORGe: Simon Mille, Stamatia Dasiopoulou, and Leo Wanner. 2019. A portable grammar-based NLG system for verbalization of structured data. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1054–1056.

Reiter&Dale's taks: Ehud Reiter and Robert Dale. 1997. *Building applied natural language generation systems.* Natural Language Engineering, 3(1):57–87.

WebNLG data: Thiago Castro Ferreira et.al.. 2020. The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results (WebNLG+ 2020). In *Proceedings of the 3rd WebNLG Workshop*, pages 55–76, Dublin, Ireland (Virtual).

Presented at INLG – Prague, Czech Republic – 11-15/09/2023

This research was funded via (i) ADAPT/DCU by the MSCA-PF-EF 2021 grant awarded for the action 101062572, and (ii) UPF by the EC-funded research and innovation programme Horizon Europe under the grant agreement number 101070278 and by the Erasmus+ programme.