Verbalizing Knowledge Graphs: Pipeline vs. End-to-End Architectures

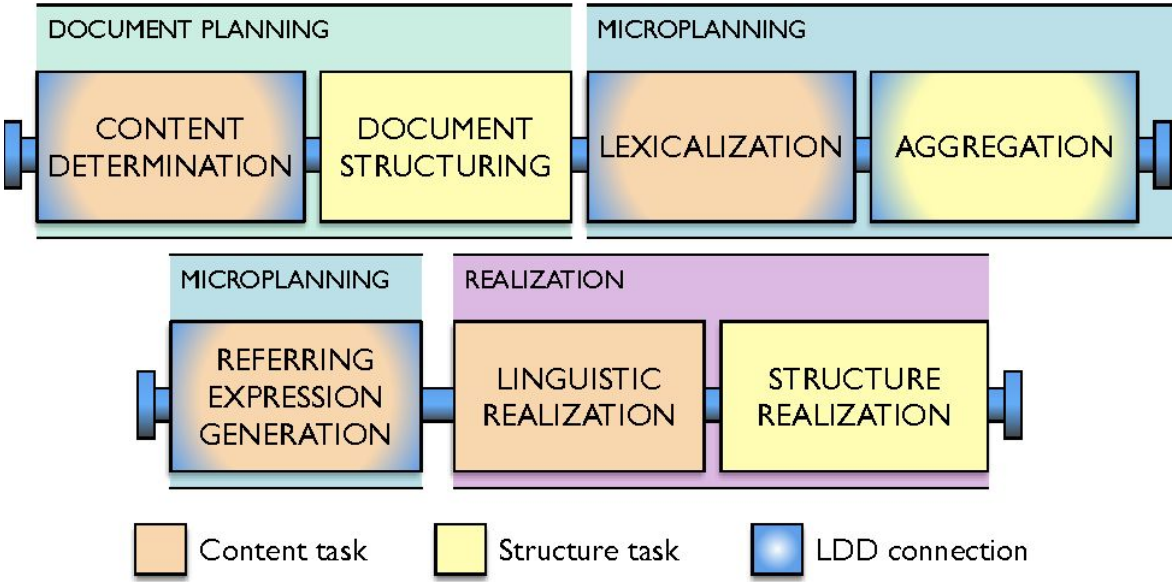
Thiago Castro Ferreira
aiXplain, inc.

Natural Language Generation

The subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other languages from some underlying non-linguistic representation of information.

(Reiter & Dale, 1997, p.1)

Classical Architecture



Pipeline

Traditionally, data-to-text models follow this architecture

Content Selection

Stimulus:

Give me information about a painting called 'Adam'

Painting	Year	Medium	Artist	Artist Nationality
Adam	1941	Watercolour	George_Pemba	South_Africa
Abaporu	1928	Oil_painting	Tarsila_do_Amaral	Brazil
Mountain_girl	1984	-	Li_Zijian	China
The_Milkmaid_(Vermeer)	1657–1658	Oil_painting	Johannes_Vermeer	Netherlands
The_Arrest_of_Pangeran_Diponegoro	1857	-	Raden_Saleh	Indonesia

Input

Painting	Year	Medium	Artist	Artist Nationality
Adam	1941	Watercolour	George_Pemba	South_Africa

Discourse Ordering

Subject	Predicate	Object
Adam	Medium	Watercolour
Adam	Artist Nationality	South_Africa
Adam	Artist	George_Pemba
Adam	Year	1941

Structuring

Paragraph	Sentence	Subject	Predicate	Object
1	1	Adam	Medium	Watercolour
1	1	Adam	Artist Nationality	South_Africa
1	1	Adam	Artist	George_Pemba
1	1	Adam	Year	1941

Lexicalization

ENTITY-1 VP[...] **be** ENTITY-2 **by** DT[...] **the** ENTITY-3 **artist** ENTITY-4 , VP[...] **execute in** ENTITY-5 .

Referring Expression Generation

Adam VP[...] be a watercolour painting on paper by DT[...] the South African artist George Pemba , VP[...] execute in 1941 .

Textual Realization

Adam is a watercolour painting on paper by the South African artist George Pemba, executed in 1941.

Give me information about a painting called 'Adam'

Content Selection

Painting	Year	Medium	Artist	Artist Nationality
Adam	1941	Watercolour	George_Pemba	South_Africa

End-to-End

Adam is a watercolour painting on paper by the South African artist George Pemba, executed in 1941.

End-to-End Approach

Aims to generate natural language from non-linguistic representation
with less explicit intermediate representations in-between.

WEATHER REPORTING

BEFORE NLG

Prior to Arria NLG, the UK Met Office would produce 60 human-authored reports a day using the data below.

00001, A' BHUIDHEANACH BHEAG,, SCOTLAND,
EUROPE,,1200,Fri,07,01,2011,0,NNW,15,13,1,VG,1021,40,27,10,1,3,N,11,9,0,VG,10,
2,55,18,6,1,6, WNW,9,6,0,VG,1022,61,13,3,1,9,WNW ,9,3,0, VG,1021,69,13, 0, 1,12, NW,9, 1,
0,0, VG, 1021, 77,12,-3,1,15,NNW,7,-1,1,VG,1021,88,9,- 5,1,18,NNW,8,5,1,VG, 1022,59,
3,2,1,21,NNW,7,10,1,VG,1020,43,16,9,1,24,N,8,12,1,VG,1019, 40,14,10,1,27,WNW,
6,6,0,VG,1020,65,10,4,1,30,WNW,5,2,0,VG,1020, 84,6,0,1,33,
NNW,4,0,0,VG,1020,90,5,-2,1,36,NNW,4,-1,0,GO,1019,92,5,-4,1,39,NNW,4,-2,1,GO,1020,93,6,
-5,1,42,N,4,5,1,VG,1022,60,7,4,1,45,E,3,VG,1020, 48,10,9,1,48, WNW,1,
11,3,VG,1019,44,6,11,1,51,WNW,4,8,2,VG,1020,6,9,6,7,1,54,WNW, 4,7,7,VG,1020,
84,5,6,1,57,NNW,4,6,9,VG,1019,90,5,5,1,60, N,3,6,9,GO,1017,91,5,5,1,63,N,3,6,
10,GO,017,92,4,5,1,66,E,3,9,3,VG,1017,76,8, 8,1,69,ESE,12,13,3,VG,1014,46,23,
10,1,72,SE,12,14,3,VG,1011,37,22,11,1,75,E,10,11,9,VG,1011,49,16,9,1,78,E,6,10,9,VG,1011,7
5,9,8,1,81,E,6,9,9,VG,1011,78, 8, 8,1,84,E,3,8,9,VG,1011,90,3, 7,1,87,ENE,1,7,3,GO,1013,96,
1013,96,2,7

AFTER NLG

Using the same data and no human intervention, the Arria NLG Engine can write 5,000 site-specific weather reports in less than a minute.

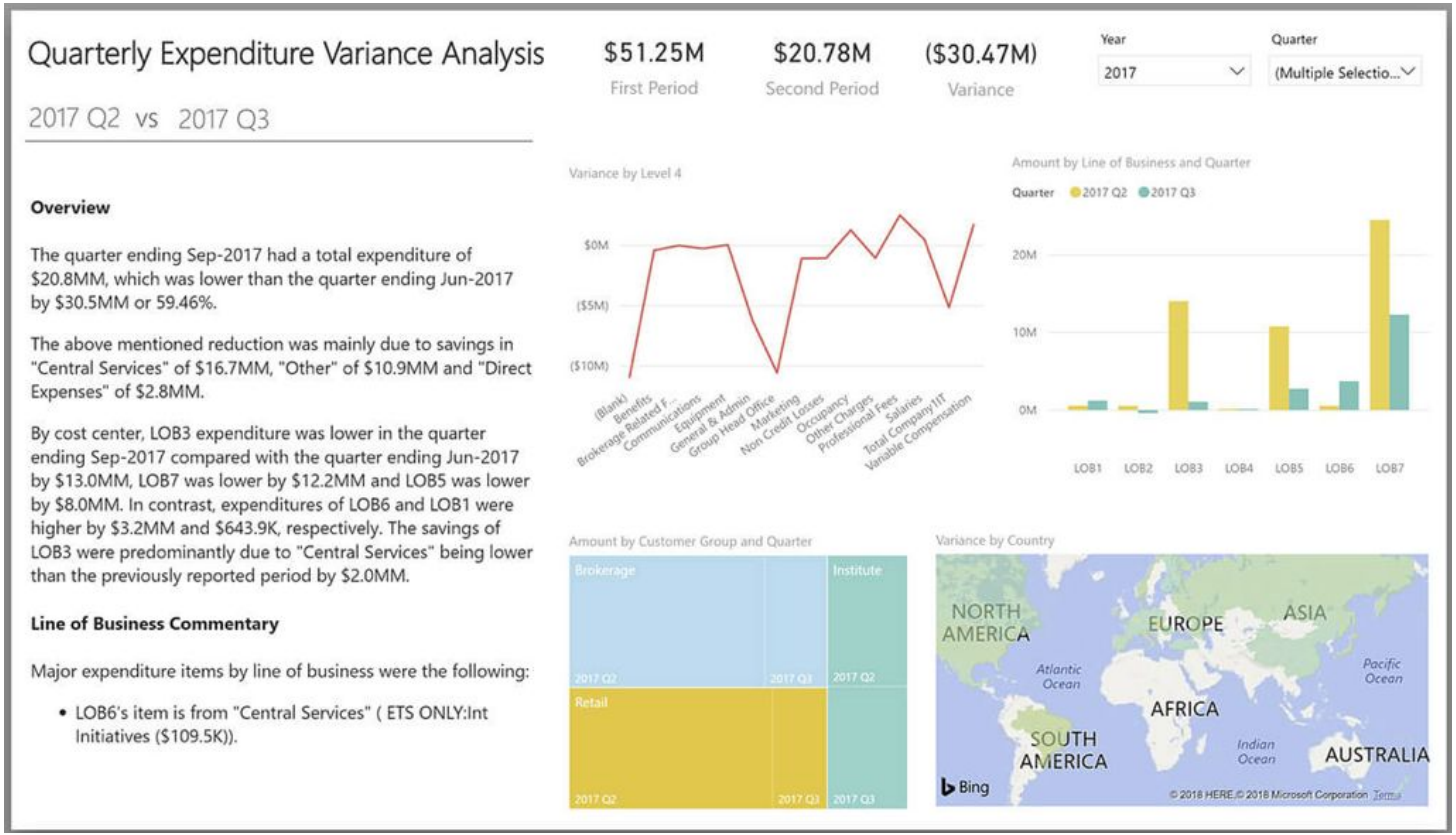
24 Hour Weather Forecast.
From 1800 on the 8th through to 1800 on the 9th.

This evening and overnight: Rain with some heavier spells spreads north over parts of Western Isles to lay from the north to the southwest of the UK by dawn tomorrow, with a chance of snow over 300m in some areas of Northeast Scotland and Central Scotland towards the early hours.

Tomorrow: The rain band lying from the north to the southwest of the UK in the early morning will move north east off shore before dying out by late afternoon. Further rain with some heavier spells reaching the northwest of the UK towards midday will move east and reach the north and west of the UK by early evening. Another area of rain will develop in some areas of Southwest England and Southern England from late afternoon onwards.

POWERED BY
ARRIA
NATURAL LANGUAGE GENERATION

NLG Applications: Financial Reports



NLG Applications: Blogging Birds

(Siddharthan et al., 2019)

NLG Applications: Verbalizers of Knowledge Graphs

dbr:Belo_Horizonte	dbo:country	dbr:Brazil
dbr:Belo_Horizonte	dbo:foundingDate	“1897-12-12”∧∧xsd:date
dbr:Belo_Horizonte	dbo:isPartOf	dbr:Minas_Gerais
dbr:Belo_Horizonte	dbo:populationTotal	“2502557”∧∧xsd:integer

Belo Horizonte is a municipality in the State of Minas Gerais in Brazil, founded on December 12, 1897. Its estimated population is 2,501,576 inhabitants.

2017 WebNLG Shared-Task

(John_E_Blaha birthDate 1942_08_26)
(John_E_Blaha birthPlace San_Antonio)
(John_E_Blaha occupation Fighter_pilot)

John E Blaha, born in San Antonio on 1942-08-26,
worked as a fighter pilot

- **Goal:** Automatically generate English summaries of sets of RDF triples
- Dataset and shared-task which straighted the bonds between the NLG and the Semantic Web research communities.

Gardent, Claire; Shimorina, Anastasia; Narayan, Shashi; Perez-Beltrachini, Laura
Creating Training Corpora for NLG Micro-Planners
In Proceedings of ACL 2017

2017 WebNLG Shared-task: Dataset

- 21,855 textual summaries to 8,372 sets of RDF triples
- Triplesets extracted from DBpedia
- Triples belonging to15 DBpedia Semantic Categories
  - **Seen:** Astronaut, University, Monument, Building, ComicsCharacter, Food, Airport, SportsTeam and WrittenWork
  - **Unseen:** CelestialBody, MeanOfTransportation, City, Athlete, Politician and Artist

2017 WebNLG Shared-Task: Participants

Rule-based Pipelines
UTilburg, Tilburg University
UIT-VNU-HCM, University of Information Technology
UPF-FORGe, Universitat Pompeu Fabra

Statistical Machine Translation-based
UTilburg-SMT, Tilburg University

End-to-End Neural Machine Translation Approaches
ADAPTCentre, ADAPT Centre Ireland
UMelbourne, University of Melbourne
UTilburg-NMT, Tilburg University
PKUWriter, Peking University
Baseline

2017 WebNLG Shared-Task: Results

Grammar (seen)	Avg	Groups
ADAPT	2.66	a
WEBNLG	2.62	b
BASELINE	2.51	c
UPF-FORGE	2.50	c
PKUWRITER	2.50	c
MELBOURNE	2.46	d
TILB-PIPELINE	2.33	e
TILB-NMT	2.23	f
TILB-SMT	2.14	g
UIT-VNU	1.51	h

Grammar (unseen)	Avg	Groups
WEBNLG	2.59	a
UPF-FORGE	2.40	b
MELBOURNE	2.27	c
PKUWRITER	2.17	d
BASELINE	2.07	e
TILB-NMT	2.01	e
TILB-PIPELINE	1.91	f
ADAPT	1.63	g
TILB-SMT	1.61	g
UIT-VNU	1.14	h

Semantics (seen)	Avg	Groups
WEBNLG	2.76	a
ADAPT	2.73	a
UPF-FORGE	2.71	a
MELBOURNE	2.48	b
TILB-SMT	2.47	b
TILB-PIPELINE	2.46	b
PKUWRITER	2.23	c
BASELINE	2.21	c
TILB-NMT	2.14	c
UIT-VNU	1.55	d

Semantics (unseen)	Avg	Groups
WEBNLG	2.79	a
UPF-FORGE	2.62	a
TILB-SMT	2.30	b
MELBOURNE	1.93	c
PKUWRITER	1.77	c
TILB-PIPELINE	1.70	cd
TILB-NMT	1.69	cd
ADAPT	1.45	de
BASELINE	1.17	e
UIT-VNU	1.16	e

Fluency (seen)	Avg	Groups
ADAPT	2.62	a
WEBNLG	2.57	b
PKUWRITER	2.44	c
BASELINE	2.40	c
UPF-FORGE	2.36	d
MELBOURNE	2.35	d
TILB-PIPELINE	2.21	e
TILB-NMT	2.10	f
TILB-SMT	2.01	g
UIT-VNU	1.50	h

Fluency (unseen)	Avg	Groups
WEBNLG	2.61	a
UPF-FORGE	2.31	b
PKUWRITER	2.14	c
MELBOURNE	2.11	c
BASELINE	1.94	d
TILB-NMT	1.85	e
TILB-PIPELINE	1.78	f
ADAPT	1.55	g
TILB-SMT	1.44	h
UIT-VNU	1.14	i

Interim Conclusion

- Divergence between automatic and human evaluation metrics
- Neural End-to-End systems seem to generate fluent, grammatical and semantic summaries for categories seen during training
- Their performance drops significantly for unseen entities, where rule-based pipelines systems are the best
- One of the first formal evidences of hallucination

Pros and Cons

Pipeline

Transparency, Easily to Control and Reusable Modules

Demands More Manual Labor

End-to-End

State-of-the-art performance in other text generation tasks (e.g., Machine Translation)

Black-box with no transparency and internal control

Experiment

Problem

Lack of an empirical comparison between both architectures

Research Question

How well does a neural pipeline architecture perform compared to a neural end-to-end one?

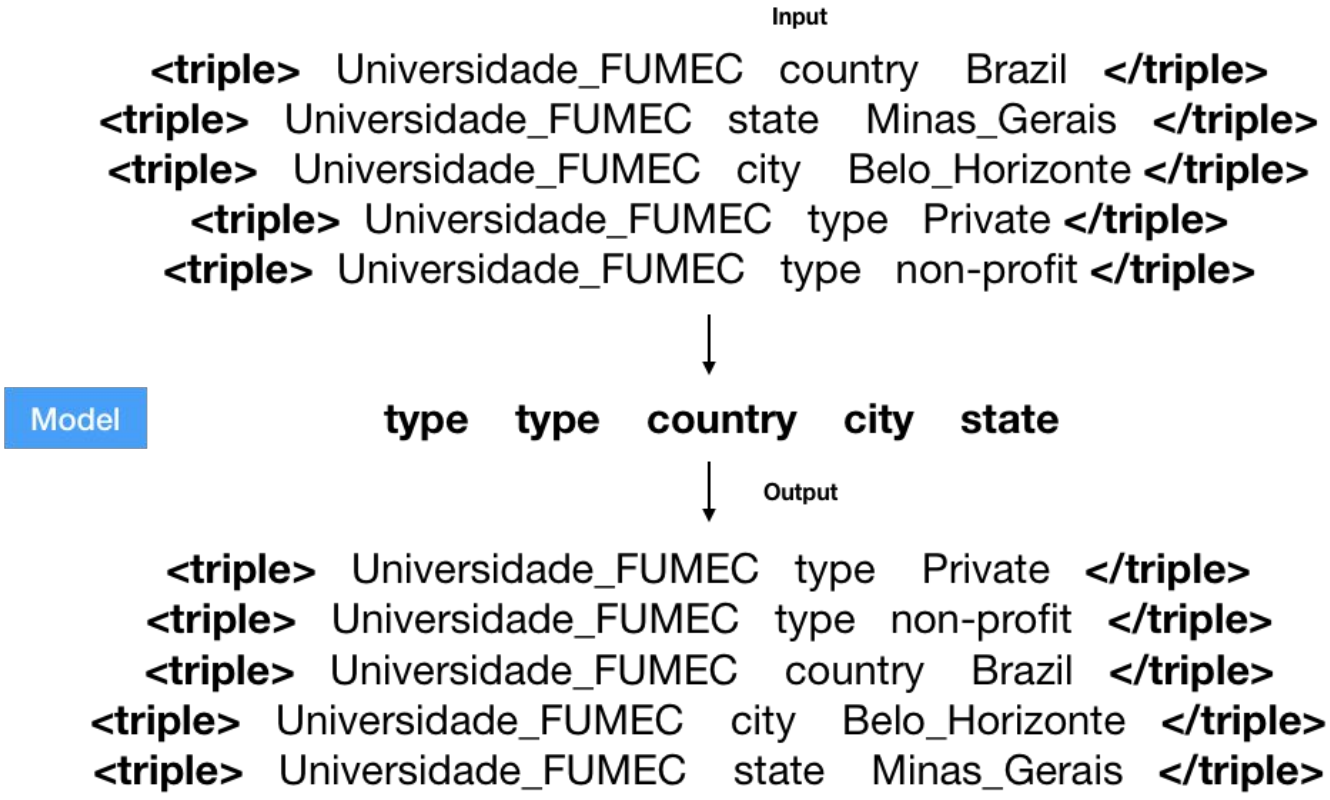
Pipeline

Step	Model
Discourse Ordering	GRU / Transformer
Text Structuring	GRU / Transformer
Lexicalization	GRU / Transformer
Referring Expression Generation	NeuralREG
Textual Realization	Rules

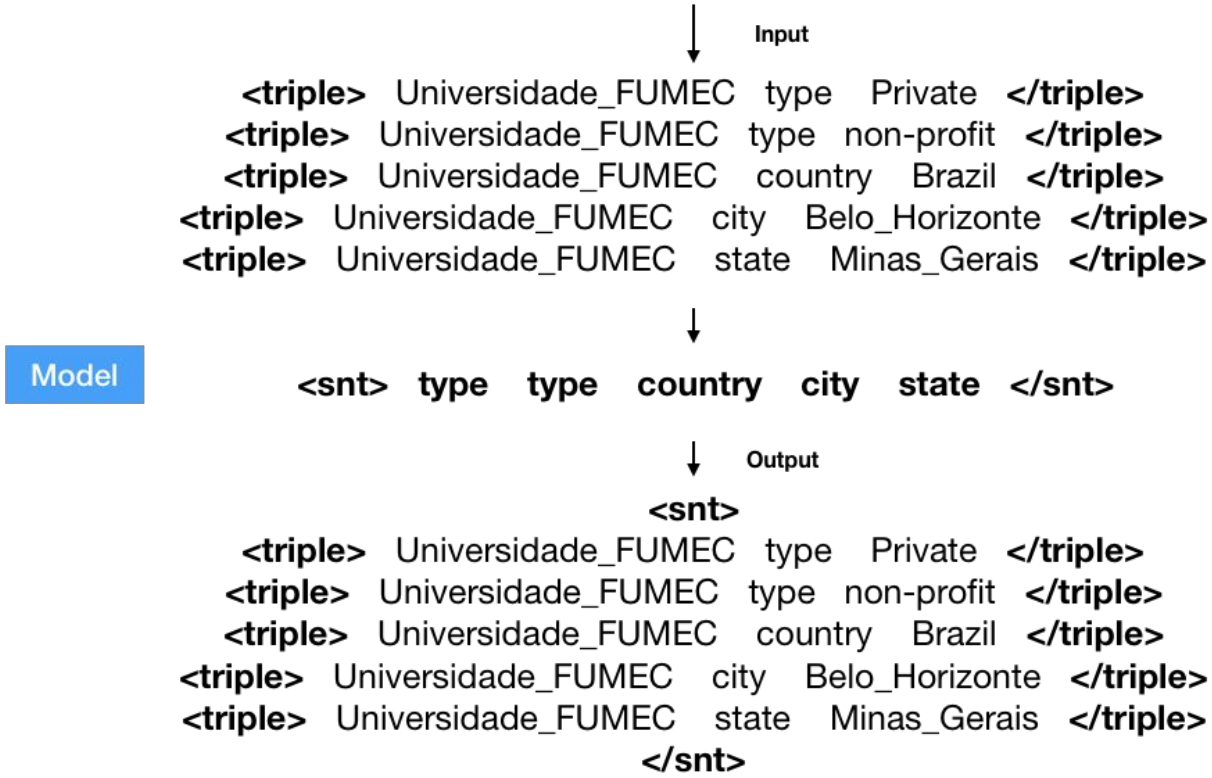
GRU
Gated-Recurrent Units

NeuralREG
(Castro Ferreira et al., 2018b)

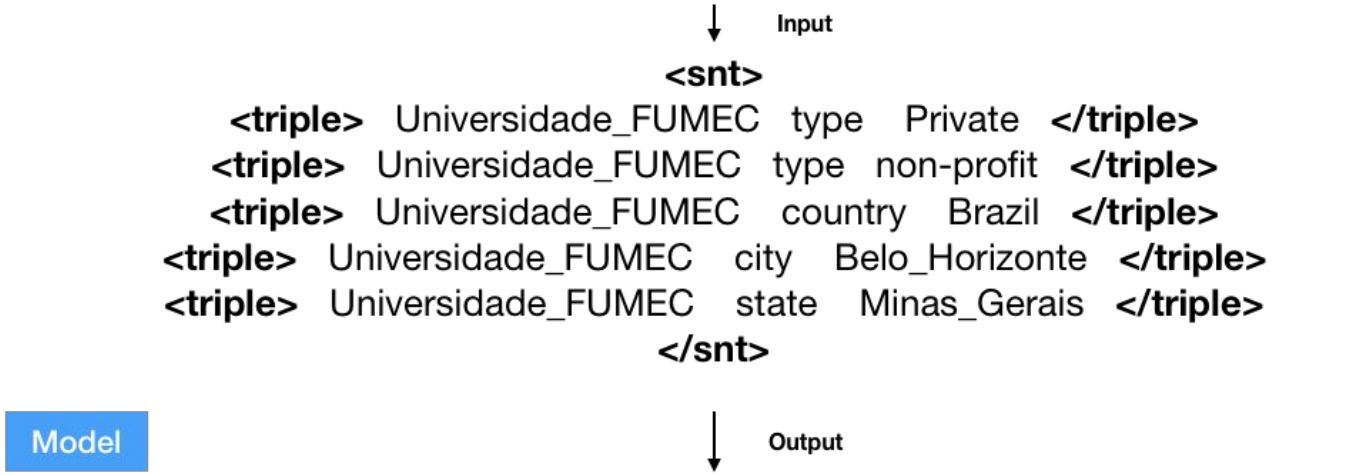
Discourse Ordering



Structuring



Lexicalization



ENTITY-1 VP[...] be DT[...] a ENTITY-2 and ENTITY-3 ENTITY-4 university in ENTITY-5 , state of ENTITY-6 .

Referring Expression Generation

NeuralREG

(Castro Ferreira et al., 2018b)

Universidade_FUMEC VP[...] be DT[...] a **Private** and **non-profit** **Brazil** university in **Belo_Horizonte** , state of **Minas_Gerais** .

NeuralREG

FUMEC University VP[...] be DT[...] a **private** and **non-profit** **Brazilian** university in **Belo Horizonte** , state of **Minas Gerais** .

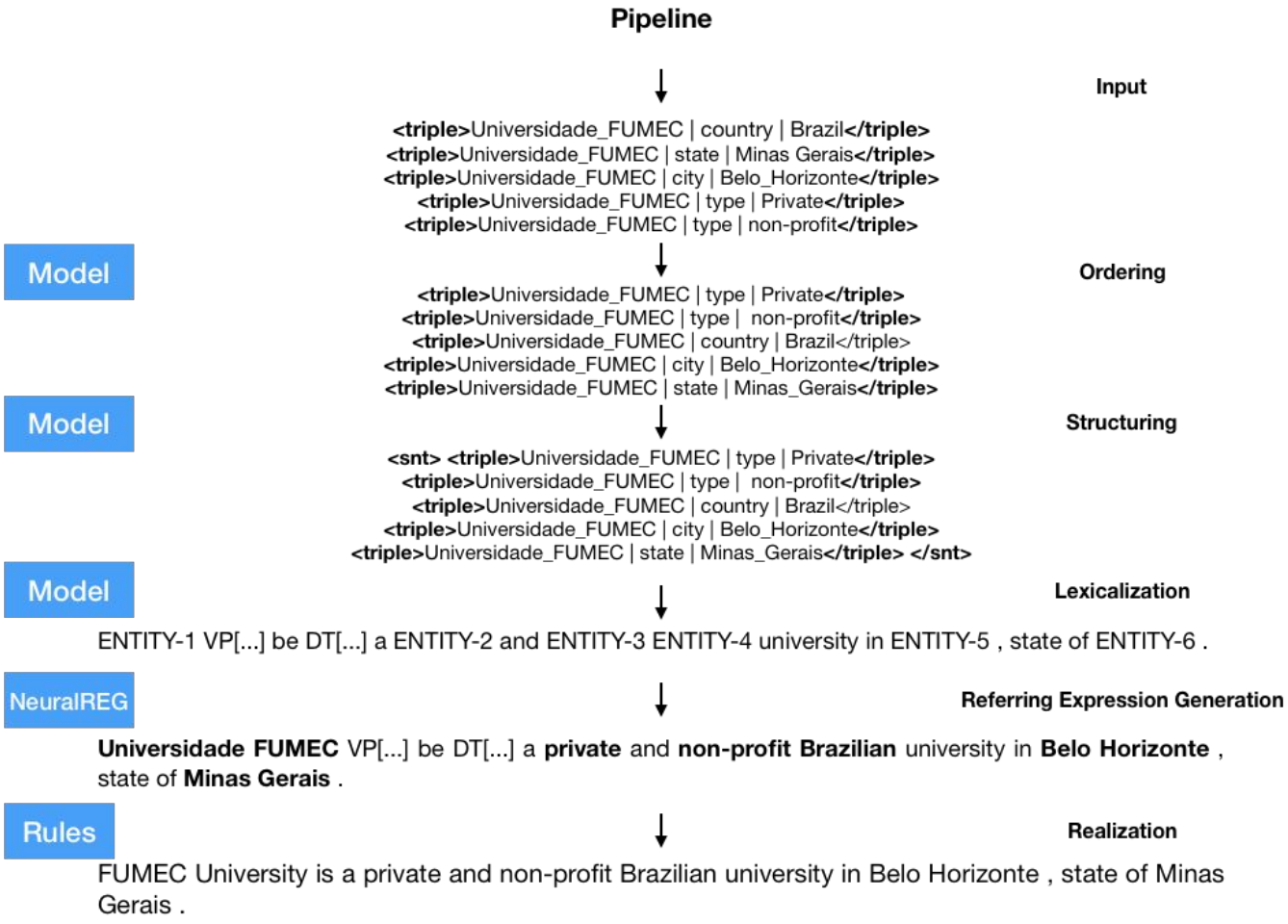
Textual Realization

Based on rules

FUMEC University **VP[...] be *DT[...] a*** private and non-profit Brazilian university in Belo Horizonte , state of Minas Gerais .

Rules

FUMEC University **is *a*** private and non-profit Brazilian university in Belo Horizonte , state of Minas Gerais .


Pipeline
Input
<triple>Universidade_FUMEC | country | Brazil</triple>
<triple>Universidade_FUMEC | state | Minas Gerais</triple>
<triple>Universidade_FUMEC | city | Belo_Horizonte</triple>
<triple>Universidade_FUMEC | type | Private</triple>
<triple>Universidade_FUMEC | type | non-profit</triple>
Model
Ordering
<triple>Universidade_FUMEC | type | Private</triple>
<triple>Universidade_FUMEC | type | non-profit</triple>
<triple>Universidade_FUMEC | country | Brazil</triple>
<triple>Universidade_FUMEC | city | Belo_Horizonte</triple>
<triple>Universidade_FUMEC | state | Minas_Gerais</triple>
Model
Structuring
<snt> <triple>Universidade_FUMEC | type | Private</triple>
<triple>Universidade_FUMEC | type | non-profit</triple>
<triple>Universidade_FUMEC | country | Brazil</triple>
<triple>Universidade_FUMEC | city | Belo_Horizonte</triple>
<triple>Universidade_FUMEC | state | Minas_Gerais</triple> </snt>
Model
Lexicalization
ENTITY-1 VP[...] be DT[...] a ENTITY-2 and ENTITY-3 ENTITY-4 university in ENTITY-5 , state of ENTITY-6 .
NeuralREG
Referring Expression Generation
Universidade FUMEC VP[...] be DT[...] a private and non-profit Brazilian university in Belo Horizonte , state of Minas Gerais .
Rules
Realization
FUMEC University is a private and non-profit Brazilian university in Belo Horizonte , state of Minas Gerais .

Baselines

Random/Majority Pipeline

Ordering, Structuring and Lexicalization

Given ***the input of predicates***, retrieve *a random/the most frequent*

discourse order → text structure → lexicalization template

in the training set

↓

Referring Expression Generation: Only Names

Universidade_FUMEC -> Universidade FUMEC

Just remove the underscore from the entity's identifier

↓

Textual Realization

Rules

End-to-End

Converts the non-linguistic representation to text in one-go, using GRUs or Transformer encoder-decoders.

<triple> Universidade_FUMEC country Brazil **</triple>**
<triple> Universidade_FUMEC state Minas_Gerais **</triple>**
<triple> Universidade_FUMEC city Belo_Horizonte **</triple>**
<triple> Universidade_FUMEC type Private **</triple>**
<triple> Universidade_FUMEC type non-profit **</triple>**

Model

FUMEC University is a private and non-profit Brazilian university in Belo Horizonte , state of Minas Gerais .

Evaluation

Automatic
BLEU and METEOR
2017 WebNLG's testset

Human
Fluency and Semantic
1-7 Likert Scale
223 pairs of WebNLG's test set

Qualitative Analysis
75 trials
Semantics and Overgeneration

Results

	Automatic		Human	
	BLEU	METEOR	Fluency	Semantic
Random	41,68	0,20	4,55	4,44
Majority	43,82	**0,33**	5,00	**5,02***
Pipe GRU	50,55	**0,33**	**5,31***	**5,21***
Pipe Transformer	**51,68**	0,32	5,03	4,87
E2E GRU	33,49	0,25	4,73	4,47
E2E Transformer	31,88	0,25	5,02	4,7

Pipeline generates more fluent and semantic texts than end-to-end

Results

	Domains seen during training				Unseen Domains			
	BLEU	METEOR	Fluency	Sem.	BLEU	METEOR	Fluency	Sem.
Random	41,72	0,27	4,79	4,73	**41,51**	-	4,07	3,86
Majority	44,79	0,41	5,25	5,41	41,13	**0,22**	4,49	4,25
Pipe GRU	55,75	**0,42**	**5,51***	**5,48***	38,55	**0,22**	**4,91***	**4,67***
Pipe Transfor	56,35	0,41	**5,53***	**5,49***	38,92	0,21	4,05	3,64
E2E GRU	**57,20**	0,41	5,40	5,21	6,25	0,09	3,45	3,03
E2E Transfor	50,79	0,39	5,38	5,15	5,88	0,09	4,32	3,81

Performance of **end-to-end** significantly drops on **unseen domains**

Example in the Unseen domain

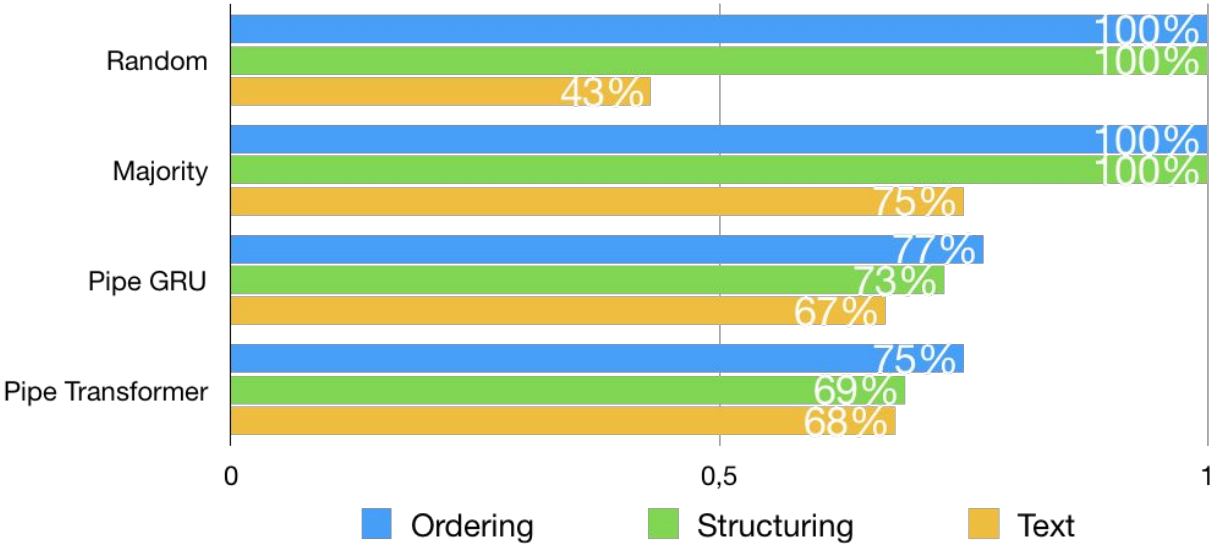
Ace_Wilder	**background**	"solo_singer"
Ace_Wilder	***birthPlace***	Sweden
Ace_Wilder	**birthYear**	1982
Ace_Wilder	***occupation***	Songwriter

↓

GRU	Ace Wilder, born in Sweden, performs as Songwriter.
Transformer	Ace Wilder (born in Sweden) was Songwriter.
E2E GRU	The test pilot who was born in Willington, who was born in New York, was born in New York and is competing in the competing in the U.S.A. The construction of the city is produced in Mandesh.
E2E Trans.	Test pilot Elliot See was born in Dallas and died in St. Louis.

Qualitative Analysis

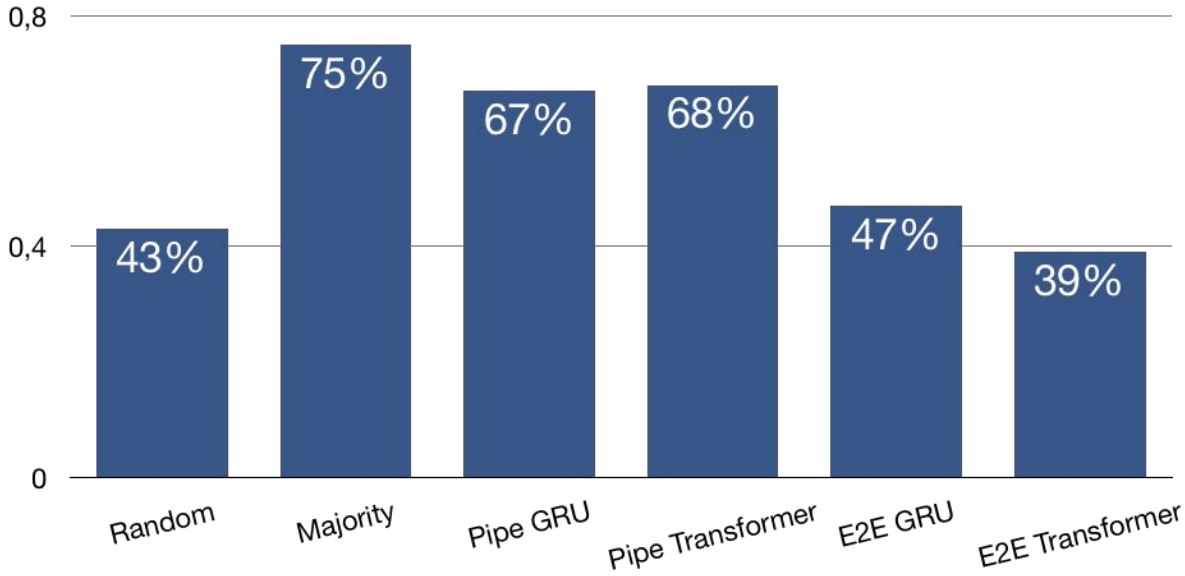
How many trials retained the content during the pipeline steps



GRU performs better at **ordering** and **structuring** the non-linguistic input, whereas **the transformer** performs better in **lexicalizing** an ordered and structured input.

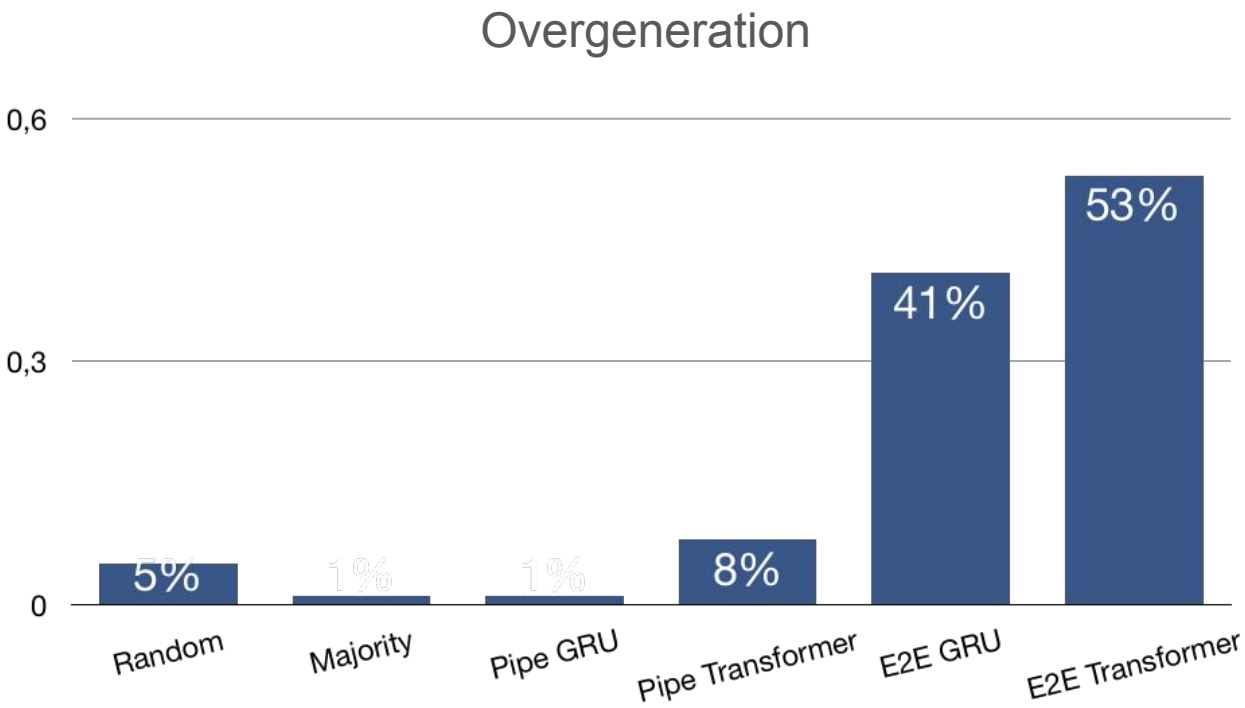
Qualitative Analysis

How many trials retained the input content in the final text



- **75% of the majority** trials retained the input content
- **Less than half of the end-to-end trials** retained the input content

Qualitative Analysis



End-to-end trials constantly contained more information than in the non-linguistic input

Interim Conclusion 2

- Pipeline generates more fluent and semantic summaries than end-to-end, and generalizes better on new domains
- End-to-End regularly works in domains seen during training, but its performance drastically drops in unseen domains, hallucinating content
- Confirms the trends in related work in favor of pipeline architectures
- Not only for the implementation of end-to-end approaches, deep learning is also a powerful disambiguation tool for particular pipeline modules

Ferreira, T.C.; van der Lee, C.; van Miltenburg, E.; Krahmer, E.
Neural data-to-text generation: A comparison between pipeline and end-to-end architectures
In Proceedings of EMNLP 2019

Problem

- Both neural pipeline and end-to-end approaches were trained from scratch
- SOTA Pre-trainer large language models were evaluated in this research

2020 WebNLG Shared-Task

Multi-task

RDF-to-text and text-to-RDF

Multilingual

English and Russian

New Semantic Categories

Seen: Company

Unseen: Scientist, Film and MusicalWork

2020 WebNLG Shared-Task: Participants

Team	Affiliation	Country	D2T		SP	
			En	Ru	En	Ru
med	Sber AI Lab	Russia	-	✓	-	-
RALI-Université de Montréal	Université de Montréal	Canada	✓	-	-	-
ORANGE-NLG	Orange Labs	France	✓	-	-	-
cuni-ufal	Charles University	Czechia	✓	✓	-	-
TGen	AIST	Japan	✓	-	-	-
bt5	Google	US	✓	✓	✓	✓
UPC-POE	Universitat Politècnica de Catalunya	Spain	✓	-	-	-
DANGNT-SGU	Saigon University	Vietnam	✓	-	-	-
Huawei Noah's Ark Lab	Huawei Noah's Ark Lab	UK	✓	✓	-	-
Amazon AI (Shanghai)	Amazon AI (Shanghai)	China	✓	-	✓	-
NILC	University of São Paulo	Brazil	✓	-	-	-
NUIG-DSI	National University of Ireland	Ireland	✓	-	-	-
CycleGT	Amazon	China	✓	-	✓	-
OSU Neural NLG	The Ohio State University	US	✓	✓	-	-
FBConvAI	Facebook	US	✓	✓	-	-

2020 WebNLG Shared-Task: Participants

Team	Pipe	E2E	Rules	Neural	Data Aug.	Delex.	T5	BART
RALI	X		X					
ORANGE-NLG		X		X	X			X
cuni-ufal		X		X				X
TGen	X			X			X	
bt5		X		X			X	
UPC-POE		X		X	X			
DANGNT-SGU	X		X			X		
Huawei	X			X		X		
Amazon Shanghai	X			X			X	
NILC		X		X				X
NUIG-DSI		X		X			X	
CycleGT		X		X			X	
OSU Neural NLG		X		X			X	
FBConvAI	X			X				X
Count	6	8	2	12	2	2	6	4

Table 3: Model Summary

Baselines

Rule-based Pipeline Systems

UPF-FORGe (2017)

FORGe2020 (Updated version)

Evaluation

Automatic

BLEU, METEOR, chrF++, TER, BERT-Score, BLEURT

Human Evaluation

Data Coverage

Does the text include descriptions of all predicates presented in the data?

Relevance

Does the text describe only such predicates (with related subjects and objects), which are found in the data?

Correctness

When describing predicates which are found in the data, does the text mention correct the objects and adequately introduces the subject for this specific predicate?

Text Structure

Is the text grammatical, well-structured, written in acceptable English language?

Fluency

Is it possible to say that the text progresses naturally, forms a coherent whole and it is easy to understand the text?

Results

	DATA COVERAGE			RELEVANCE			CORRECTNESS			TEXT STRUCTURE			FLUENCY		
TEAM NAME	**Rank**	**Avg. Z**	**Avg. Raw**	**Rank**	**Avg. Z**	**Avg. Raw**	**Rank**	**Avg. Z**	**Avg. Raw**	**Rank**	**Avg. Z**	**Avg. Raw**	**Rank**	**Avg. Z**	**Avg. Raw**
FBConvAI *	2	0.151	93.169	2	0.117	93.898	1	0.206	92.700	1	0.319	93.089	1	0.327	90.837
AmazonAI (Shanghai)	1	0.222	94.393	1	0.214	95.196	1	0.248	93.531	1	0.305	92.951	1	0.326	90.286
OSU Neural NLG	1	0.235	95.123	1	0.163	94.615	1	0.224	93.409	1	0.289	92.438	1	0.323	90.066
WebNLG-2020-REF	1	0.251	95.442	1	0.139	94.392	1	0.256	94.149	1	0.254	92.105	1	0.279	89.846
NUIG-DSI	2	0.116	92.063	1	0.161	94.061	1	0.189	92.053	1	0.258	91.588	2	0.233	88.898
bt5	2	0.161	93.836	1	0.184	95.220	1	0.224	93.583	1	0.236	91.914	2	0.218	88.688
cuni-ufal	2	0.155	93.291	1	0.164	94.555	1	0.161	91.587	1	0.208	90.752	2	0.185	87.642
TGen	3	-0.075	88.176	1	0.132	92.640	2	0.074	88.626	1	0.168	89.041	2	0.182	86.163
CycleGT	3	0.023	91.231	1	0.125	93.370	2	0.071	89.846	2	0.045	87.879	3	0.072	84.820
Baseline-FORGE2020	1	0.170	92.892	1	0.161	93.784	1	0.190	91.794	2	0.039	87.400	3	0.011	82.430
Baseline-FORGE2017	2	0.127	92.066	2	0.113	92.588	2	0.13	90.138	2	-0.064	85.737	4	-0.143	80.941
DANGNT-SGU	1	0.259	95.315	1	0.185	94.856	1	0.179	92.489	3	-0.203	83.501	4	-0.161	78.594
RALI-Université de Montréal	1	0.272	95.204	1	0.171	94.810	1	0.163	92.128	3	-0.285	81.835	4	-0.241	77.759
ORANGE-NLG	5	-0.554	79.959	4	-0.710	79.887	4	-0.668	74.977	3	-0.338	80.462	5	-0.332	75.675
Huawei Noah's Ark Lab	4	-0.310	84.743	3	-0.425	85.265	3	-0.389	80.760	3	-0.373	80.219	5	-0.369	75.205
NILC	4	-0.477	81.605	3	-0.499	83.522	3	-0.589	76.702	3	-0.402	80.463	5	-0.408	74.851
UPC-POE	6	-0.782	75.845	4	-0.531	82.051	4	-0.701	74.374	4	-0.456	78.503	5	-0.508	72.280

Interim Conclusion 3

- Pre-trained LLMs obtained parity with human references in fluency and text structure
- Although with less modules, neural pipeline models outperform end-to-end ones in terms of adequacy
  - AmazonAI (Shanghai) and FBConvAI keep a separate discourse ordering module
  - Their results are similar to ruled-based approaches, SOTA in the previous version of the shared-task
  - Modules like text structuring, lexicalization, referring expression generation and textual realization did not seem necessary in WebNLG's context

Ferreira, T.C.; Gardent, Claire; Ilinykh, Nikolai; van der Lee, Chris; Mille, Simon; Moussallem, Diego; Shimorina, Anastasia
The 2020 Bilingual, Bi-Directional WebNLG+ Shared Task Overview and Evaluation Results (WebNLG+ 2020)
In Proceedings of WebNLG+ 2020

Experiment

Problem

Lack of an systematic comparison of pre-trained large language models verbalizers

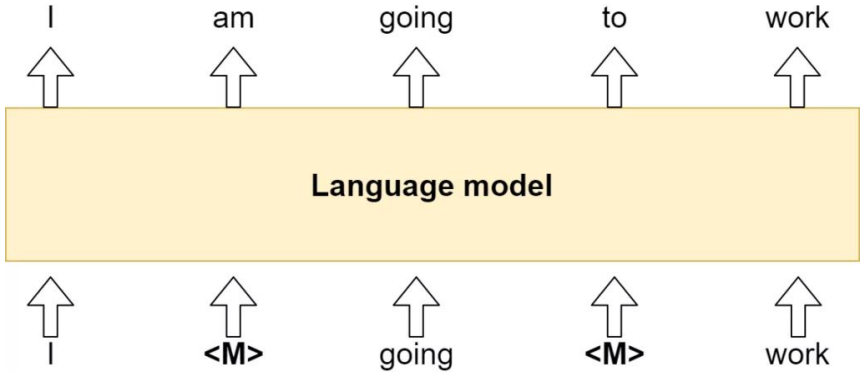
Research Question

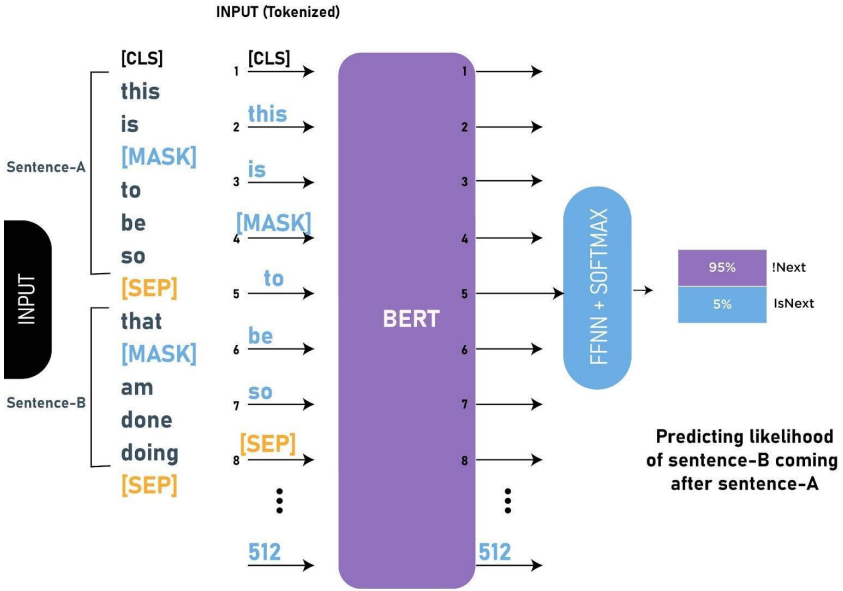
Where do a LLM generation model fail?

Evaluated Approaches

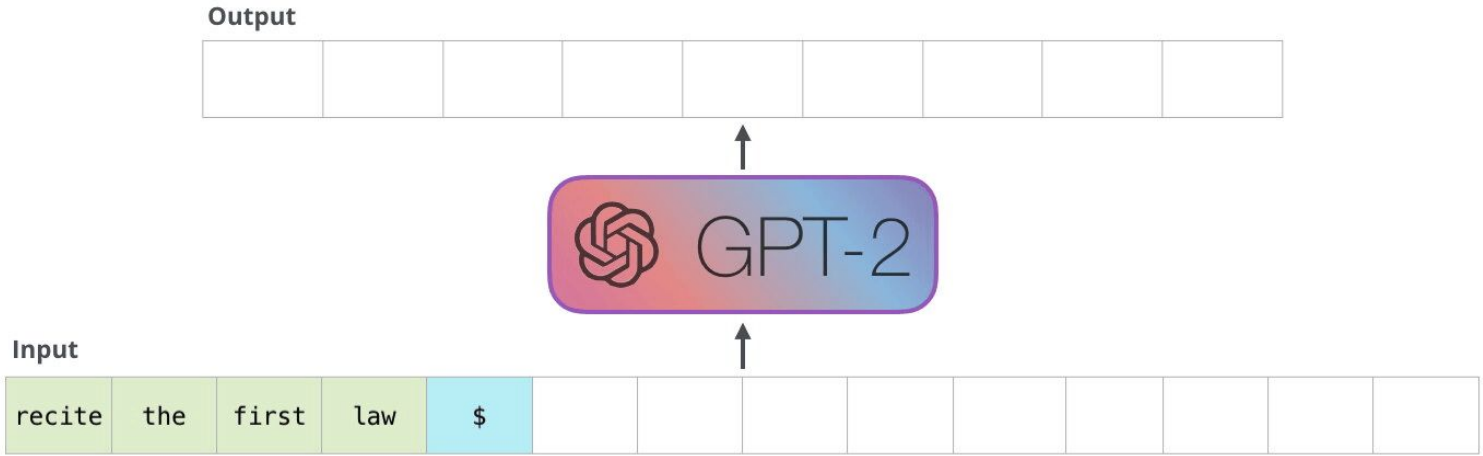
- BERTimbau
- GPorTuguese-2
- mT5
- mBART-50

BERTimbau (Portuguese version of BERT)

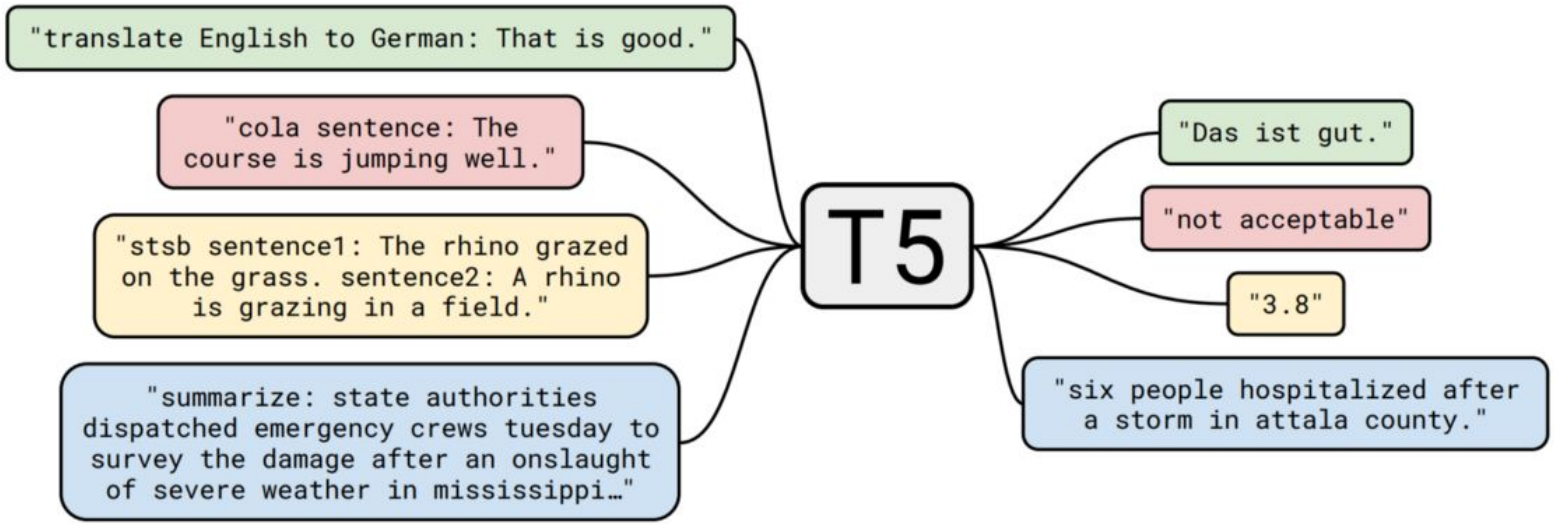




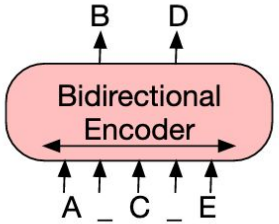
GPorTuguese-2 (Portuguese version of GPT-2)



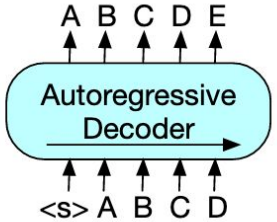
mT5


"translate English to German: That is good."
"cola sentence: The course is jumping well."
"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."
"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"
T5
"Das ist gut."
"not acceptable"
"3.8"
"six people hospitalized after a storm in attala county."

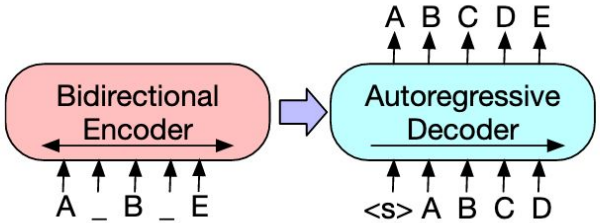
mBART-50



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with a mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Training Data

3,196 instance

@DaMataReporter

Daily Deforestation Alerts

Daily Fire Alerts

Monthly Deforestation Reports

@CoronaReporter

Daily COVID-19 Reports

Training Data: Format

Input

[COVID19] [INTENTS] TOTAL_DEATHS(deaths="99702") [HISTORY] Brazil registers 49,502 new cases of #COVID19, totaling 2,967,064 cases in total and an increase of 1.70% compared to the previous day.

Output

The country reaches the mark of 99,702 deaths for #COVID19.

Evaluation

(Thomson and Reiter, 2020)

Performed by 3 annotators in 130 test trials per model

1. Incorrect Number
2. Incorrect Named Entity
3. Incorrect Word
4. Context Error
5. Not checkable
6. Fluency
7. Others

Evaluation: Incorrect Number

Gold-Standard

Today, the Atlantic Forest registered 5 fires alerts and was the most affected biome.

Generated Text

Today, the Atlantic Forest registered 566 fires alerts and was the most affected biome.

Evaluation: Incorrect Named Entity

Gold-Standard

São Félix do Xingu totaled 70.26 km² of deforested area in September.

Generated Text

The most affected municipality was Poconé, with 70.26 km² of deforested area.

Evaluation: Context Error

Gold-Standard

The National Institute for Space Research (INPE) alerted the deforestation of 3.33 km² in the municipality of Manicore / Amazonas on July 6, 2020 - Manicore already had deforestation alerts in 6 days of this month, totaling 16.07 km² of deforestation.

Generated Text

On July 6, 2020, the National Institute for Space Research (INPE) alerted the deforestation of 3.3 km2 in the city of Manicore/AM.

Evaluation: Not Checkable

Gold-Standard

A total of 185,687 confirmed deaths by #COVID19 are recorded in the country.

Generated Text

The total number of deaths comes to 185,687, a daily increase of 6% in the total number of infected.

Evaluation: Fluency

Gold-Standard

Deaths from #COVID19 in Brazil reach 290,525, with 2,730 new deaths, representing a daily increase of 0.95%.

Generated Text

The National Institute for Space Research (INPE), the Brazil registers 283513 deaths of #COVID19 in Brazil.

Evaluation: Others

Gold-Standard

A total of 185,687 confirmed deaths by #COVID19 are recorded in the country.

Generated Text

A total of 185 , 687 confirmed deaths by # COVID19 are recorded in the country.

Preliminary Results

Model	Incorrect Number	Incorrect Named Entity	Incorrect Word	Context Error	Not Checkable	Fluency	Other
Gold	0.8%	0.8%	1.5%	0.8%	0.00%	0.8%	1.5%
BERTimbau	56.2%	16.2%	3.1%	**1.5%**	0.8%	**0.8%**	46.2%
GPorTuguese-2	**22.3%**	6.2%	**2.3%**	9.2%	2.3%	1.5%	**3.8%**
mT5	31.5%	**0.8%**	**2.3%**	2.3%	**0.00%**	1.5%	4.6%
mBART	26.9%	1.5%	3.1%	9.2%	2.3%	1.5%	4.6%

Preliminary Results

Model	Incorrect Number	Incorrect Named Entity	Incorrect Word	Context Error	Not Checkable	Fluency	Other
Gold	0.8%	0.8%	1.5%	0.8%	0.00%	0.8%	1.5%
BERTimbau	56.2%	16.2%	3.1%	**1.5%**	0.8%	**0.8%**	46.2%
GPorTuguese-2	**22.3%**	6.2%	**2.3%**	9.2%	2.3%	1.5%	**3.8%**
mT5	31.5%	**0.8%**	**2.3%**	2.3%	**0.00%**	1.5%	4.6%
mBART	26.9%	1.5%	3.1%	9.2%	2.3%	1.5%	4.6%

Produced texts are fluent and does not contain information which is not in the input

Preliminary Results

Model	Incorrect Number	Incorrect Named Entity	Incorrect Word	Context Error	Not Checkable	Fluency	Other
Gold	0.8%	0.8%	1.5%	0.8%	0.00%	0.8%	1.5%
BERTimbau	56.2%	16.2%	3.1%	**1.5%**	0.8%	**0.8%**	46.2%
GPorTuguese-2	**22.3%**	6.2%	**2.3%**	9.2%	2.3%	1.5%	**3.8%**
mT5	31.5%	**0.8%**	**2.3%**	2.3%	**0.00%**	1.5%	4.6%
mBART	26.9%	1.5%	3.1%	9.2%	2.3%	1.5%	4.6%

Low rate of orthographic mistakes, although above gold-standard

Preliminary Results

Model	Incorrect Number	Incorrect Named Entity	Incorrect Word	Context Error	Not Checkable	Fluency	Other
Gold	0.8%	0.8%	1.5%	0.8%	0.00%	0.8%	1.5%
BERTimbau	56.2%	16.2%	3.1%	**1.5%**	0.8%	**0.8%**	46.2%
GPorTuguese-2	**22.3%**	6.2%	**2.3%**	9.2%	2.3%	1.5%	**3.8%**
mT5	31.5%	**0.8%**	**2.3%**	2.3%	**0.00%**	1.5%	4.6%
mBART	26.9%	1.5%	3.1%	9.2%	2.3%	1.5%	4.6%

Low rate of named entity and context errors
Higher variation between the models

Preliminary Results

Model	Incorrect Number	Incorrect Named Entity	Incorrect Word	Context Error	Not Checkable	Fluency	Other
Gold	0.8%	0.8%	1.5%	0.8%	0.00%	0.8%	1.5%
BERTimbau	56.2%	16.2%	3.1%	**1.5%**	0.8%	**0.8%**	46.2%
GPorTuguese-2	**22.3%**	6.2%	**2.3%**	9.2%	2.3%	1.5%	**3.8%**
mT5	31.5%	**0.8%**	**2.3%**	2.3%	**0.00%**	1.5%	4.6%
mBART	26.9%	1.5%	3.1%	9.2%	2.3%	1.5%	4.6%

High rate of number errors

Conclusion

Towards End-to-End

SOTA NLG approaches have had less explicit intermediate representations along the time

High Fluency, Some Problems in Adequacy

SOTA End-to-End Large Language Models can generate fluent English text, but with some errors (e.g., numerical references)

Thank you :-)

Questions?