Robust Data-to-text Generation with Pretrained Language Models

Ondřej Dušek
collaboration with Zdeněk Kasner and Ioannis Konstas

Prague Computer Science Seminar

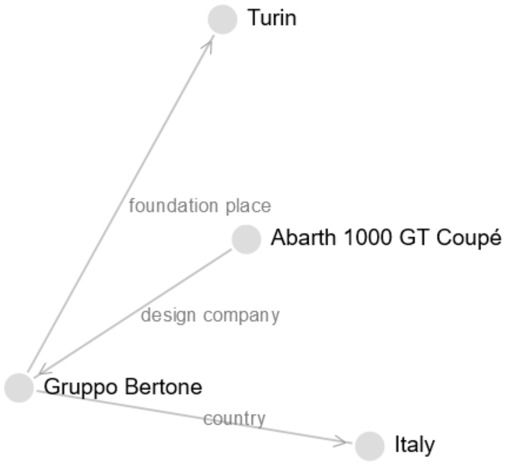
9.2.2023

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

European Research Council
Established by the European Commission

Data-to-text Generation

- **data-to-text NLG** = verbalizing structured outputs
  - RDF triples (=2 entities & relation), tables, dialogue acts … → text



Abarth 1000 GT Coupé | design company | Gruppo Bertone
Gruppo Bertone | foundation place | Turin
Gruppo Bertone | country | Italy

→ NLG →

Gruppo Bertone, of Turin Italy, designed the Abarth 1000 GT Coupe.

- main usage:
  - reports based on data (weather, sports…)
  - dialogue systems (Siri/Google/Alexa…)

Team	Win	Loss	Pts	…
Mavericks	31	41	86	
Raptors	44	29	94	

Player	AS	RB	PT	…
Patrick Patterson	1	5	14	
Delon Wright	4	3	8	
…				

→

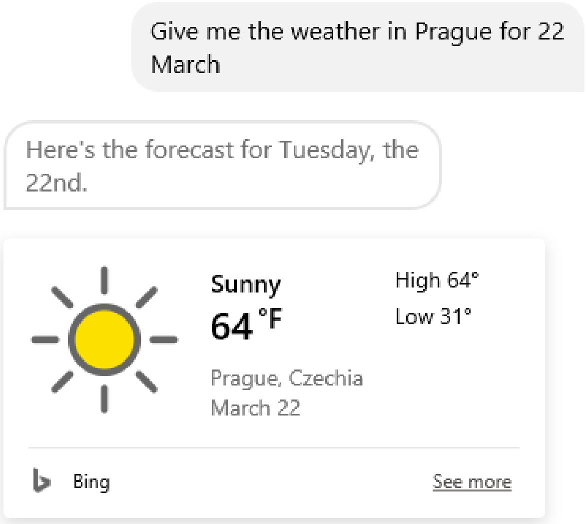
- *The Toronto Raptors, which were leading at halftime by 10 points (54-44), defeated the Dallas Mavericks by 8 points (94-86).*

…

- *Patrick Patterson provided 14 points on 5/6 shooting, 5 rebounds, 3 defensive rebounds, 2 offensive rebounds and 1 assist.*

…

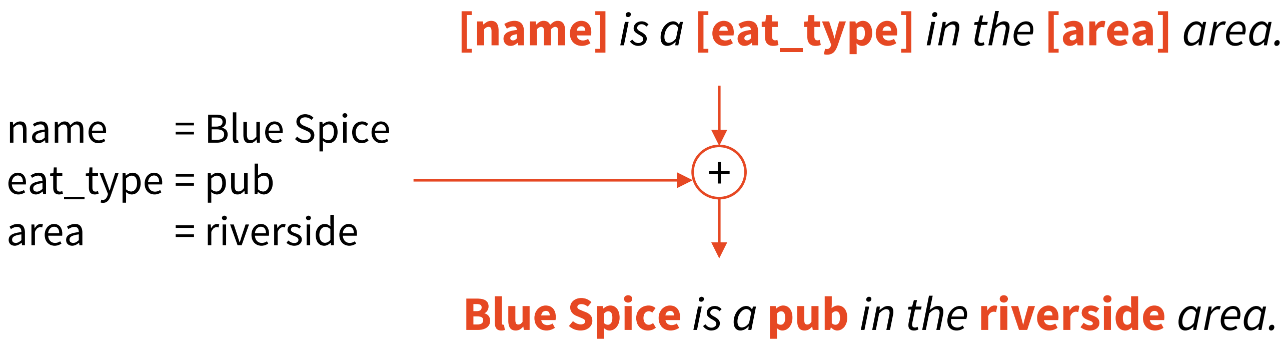
(Kasner et al., 2021) https://aclanthology.org/2021.inlg-1.25



Cortana

Neural NLG vs. older methods

- Older methods:
  - **templates** – fill in blanks
    - most commercial systems still!
    - safe, tried & tested
    - needs handcrafting
  - grammars & older statistical
    - experimental, clunky, pipelines



- Neural models:
  - 1 step, **end-to-end**
  - **Train** fully from input-output pairs (no additional rules etc.)
  - Much more **fluent** outputs
  - Needs more training data (~10k range, 10x more than before)
  - Opaque & has **no guarantees on accuracy**

Accuracy in NLG

- **NLG semantic accuracy** (fidelity) = input-output correspondence
- Basic error types:
  - **hallucination** = output not grounded in input
    - conflicting with input / unrelated to it
  - **omission** = input not verbalized

omitted

Blue Spice | eat_type | pub
Blue Spice | area | riverside

NLG

You can bring your kids to Blue Spice in the riverside area.

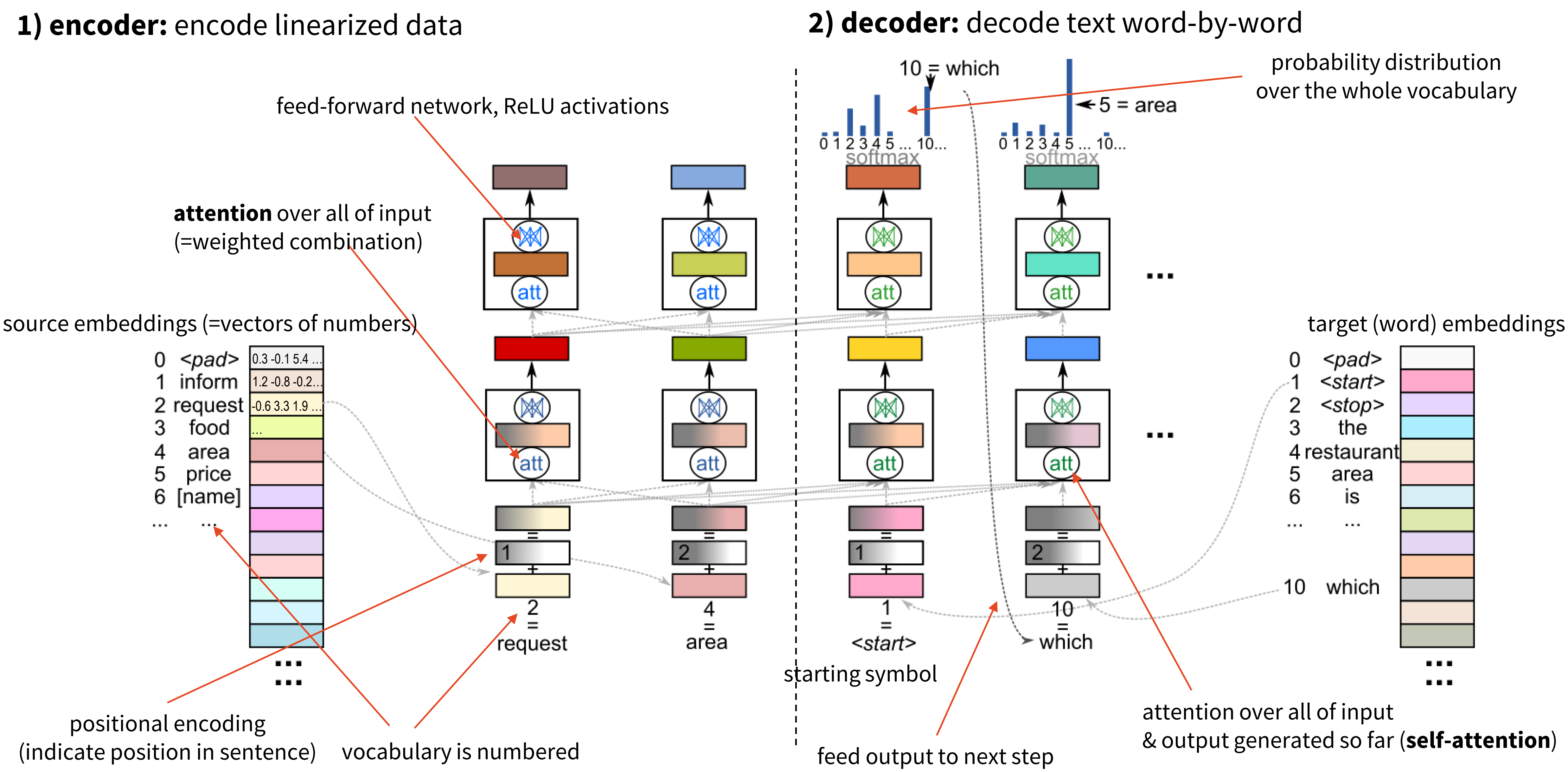
hallucinated
(not grounded)

- Approx. measure: logical entailment (NLI)
  - output entailed by data & vice-versa, neural models available (BART-NLI)

(Dušek & Kasner, 2020) https://aclanthology.org/2020.inlg-1.19/

Neural NLG: Transformer Models

(Vaswani et al., 2017) http://arxiv.org/abs/1706.03762



Neural NLG: Training

- Trained to produce sentences from data
  - replicate exact word at each position
- **Supervised** learning
  - initialize model with random parameters
  - didn't hit the right word → incur **loss**, update parameters

Blue Spice | price | expensive → NLG → *Blue Spice is expensive*

cheap

pricey

in the expensive price range

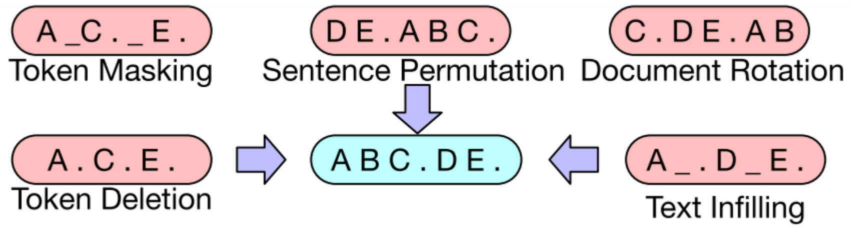
reference:

Blue Spice is expensive

- Very **low level**, no concept of sentence / text / aim

Neural NLG: Pretraining + Finetuning/Prompting

- Pretrained language models (PLMs):
  1. **Pretrain** a model on huge data (**self-supervised**, language-based tasks)
     - text-to-text (~ editing)
     - autoencoding & denoising
  2. **Fine-tune** for your own task
     on your smaller data (**supervised**)
     - same as (↑), but much better starting point
  - Models free for download (https://huggingface.co/)
     - BERT/RoBERTa, GPT-2, BART, T5… ~ 100k-1B parameters



(Lewis et al., 2020)
https://www.aclweb.org/anthology/2020.acl-main.703

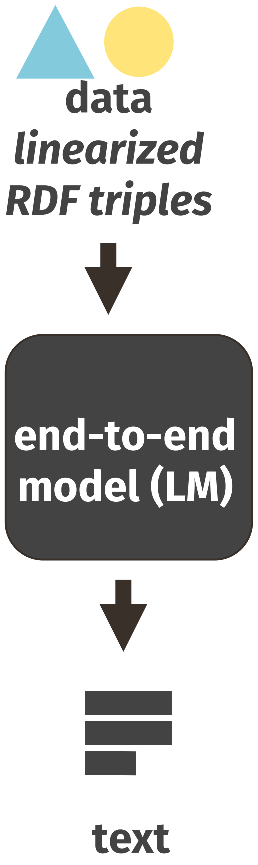
- Large language models (LLMs): **Pretrain** & **prompt**
  - 10-100B parameters, hard to run in-house (OPT, BLOOM) or not free (GPT-3, ChatGPT, LaMDa)
  - some have better pretraining (reinforcement learning)
  - feed in 1-5 examples / ask question: no need to finetune

https://jlibovicky.github.io/2023/02/07/Otazky-a-odpovedi-o-ChatGPT-a-jazykovych-modelech.html

End-to-end NLG with a Pretrained LM

- Use a pretrained LM
  - e.g. (m)BART (GPT-2, T5… ~ 100M-1B params)
- Linearize data
  - concatenate, tokenize data
- Finetune PLM
  - direct data-text mapping: black box
  - needs domain-specific data
    - scarce (~10k max)
    - noisy (crowdsourced)
- Alternative: prompt LLM
  - little/no data needed, but even less controllable

Arrabiata sauce | country | Italy ▸ Italy | capital | Rome



Arrabiata sauce is found in Italy where capital city is Rome.

NLG with a pretrained LM: Results

(Kasner & Dušek, 2020)
https://aclanthology.org/2020.webnlg-1.20/

Good

- Generally fluent and accurate
- Robust on input perturbations
- Can be multilingual

in: *Bakewell tart | ingredient | Frangipane*

out: Франжипан - один из ингредиентов тарта Бейквелл.
(=Frangipane is one of the ingredients of the Bakewell tart.)

Bad

- Fails to generalize
  - factual or grammar errors
  - specifically on unseen relations
- Hallucinations
  - connects unrelated data

in: *Ciudad_Ayala | populationMetro | 1777539*

out: The population metro of Ciudad Ayala is 1777539.

not seen in training data

in: *Nurhan Atasoy | birth date | 1934-01-01 ▸*
Nurhan Atasoy | residence | Istanbul ▸
Nurhan Atasoy | nationality | Turkish people

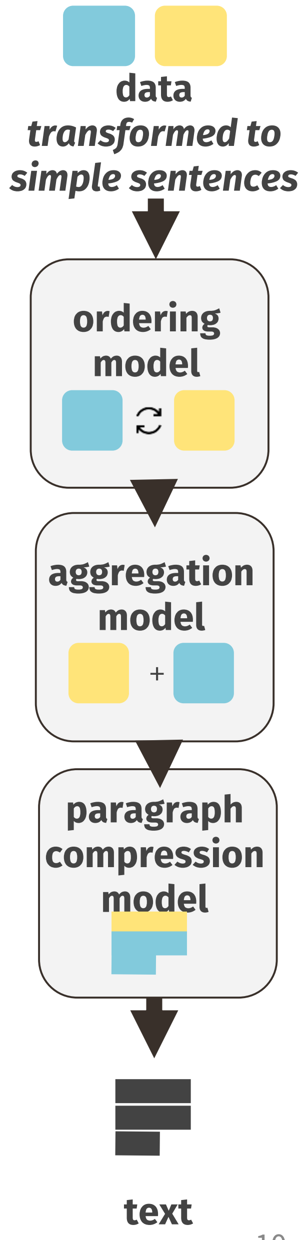
out: Nurhan Atasoy was born on January 1, 1934 in Istanbul and is a Turkish national.

residence, not birthplace!

Fuse & Rephrase Pipeline: LMs to edit only

(Kasner & Dušek, 2022)
https://aclanthology.org/2022.acl-long.271/

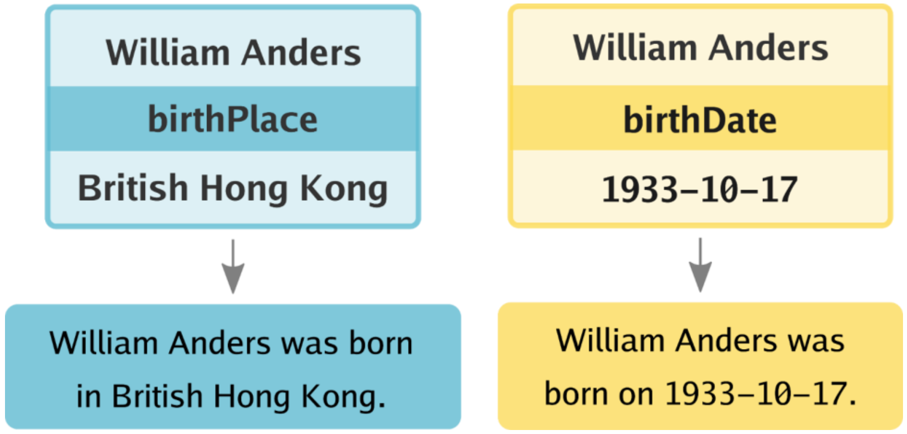
- Represent input triples by templates
  - handcrafted preprocessing step
- Neural LMs to **fuse & rephrase:**
  - All text-to-text steps (=editing only)
  1) **order** (put related stuff together)
  2) **aggregate** (into sentences)
  3) **compress** (produce shorter sentences)
- Less space for semantic errors
  - Only use LMs for what they're good at – fluency
- Can use large general-domain data (~1M+)
- Works **zero-shot** – needs no in-domain data (just the templates)



Templates

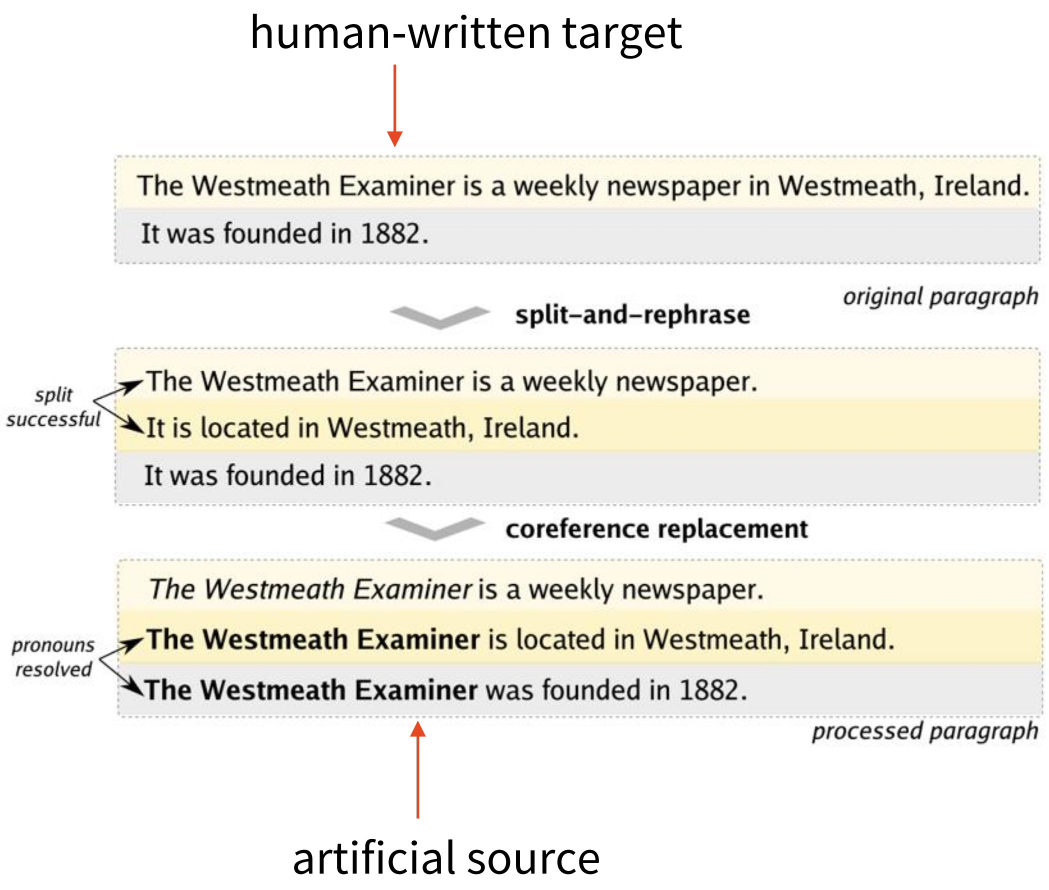
- 1 template per relation in data
  - Not so many needed (usually)
    - 354 for WebNLG DBPedia knowledge
    - 8 for E2E restaurants
  - Entities inserted verbatim
- Guaranteed accurate
- No need for high fluency
  - Some entities may need adjusting
  - LMs in the pipeline should deal with that

dataset	predicate	template
WebNLG	instrument	<s> *plays* <o>.
	countryOrigin	<s> *comes from* <o>.
	width	<s> *is* <o> *wide.*
E2E	eatType	<s> *is a* <o>.
	food	<s> *serves* <o> *food.*
	area	<s> *is in the* <o>.

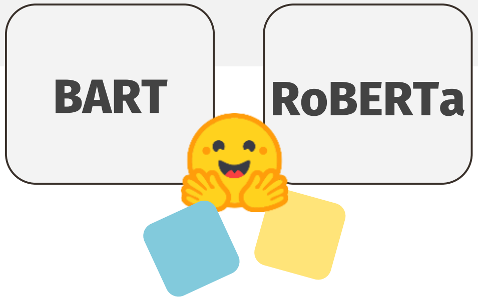


WikiFluent Corpus

- Wikipedia 1st paragraphs
  - human-written sentences as targets
  - creating artificial source data resembling single-triple templates
- Data creation process:
  1) split sentences (split & rephrase LM)
  2) replace pronouns
  3) randomize order
  4) opt. filter by logical entailment (NLI LM)
- much bigger than in-domain data (~1M sentences)

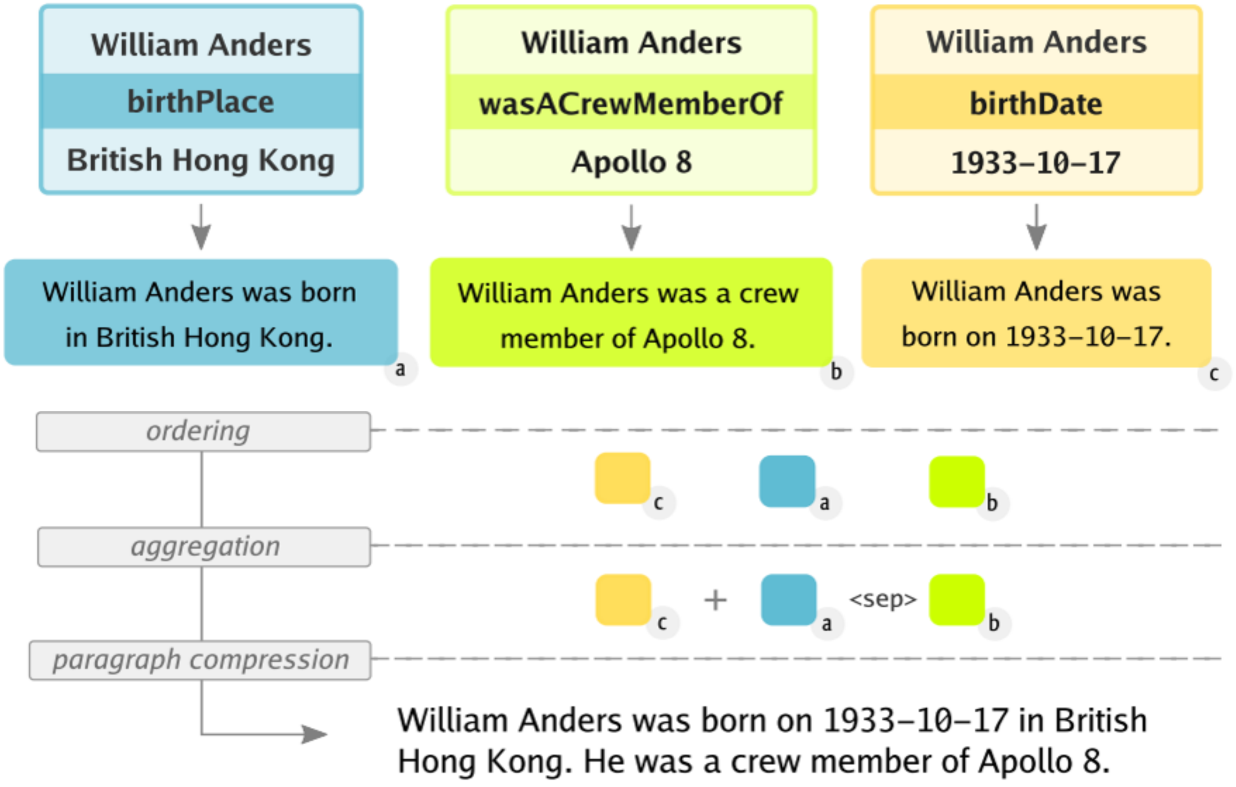


Pipeline modules



1) **Ordering**
   - BART LM with a pointer network
2) **Aggregation**
   - RoBERTa LM + token classification
   - 0/1: same/other sentence
3) **Paragraph compression**
   - BART LM – generation
   - close to pretraining tasks

- All trained on WikiFluent
  - 1M general-domain data
  - no in-domain data



Templates + Neural Fuse & Rephrase

- Good accuracy
  - perfect for simpler data (E2E – restaurants)
  - worse for complex data (WebNLG – DBPedia)
  - still merging unrelated facts on WebNLG
- Slightly lower fluency (~older neural systems)
  - still much better than templates
- 3-stage setup better than 1-stage (~end-to-end edit)
- Manual templates are cumbersome (→→)

E2E	BLEU	Omission/ #facts	Hallucination /#examples
Older neural	40.73	0.016	0.083
Templates	24.19	0.000	0.000
Ours 1-stage	30.81	0.009	0.122
Ours 3-stage	36.04	0.001	0.001

WebNLG	BLEU	Omission/ #facts	Hallucination /#examples
Rule-based	38.65	0.075	0.101
Older neural	45.13	0.237	0.202
Templates	37.18	0.000	0.000
Ours 1-stage	39.08	0.071	0.204
Ours 3-stage	42.92	0.051	0.148

Example outputs

E2E

input: *The Cricketers | eatType | restaurant ► The Cricketers | near | All Bar One ► The Cricketers | priceRange | cheap ► The Cricketers | food | Chinese ► The Cricketers | customerRating | average ► The Cricketers | familyFriendly | yes*

templates: *The Cricketers is a restaurant. The Cricketers is near All Bar One. The Cricketers has cheap price range. The Cricketers serves Chinese food. The Cricketers has average customer rating. The Cricketers is family-friendly.*

output: The Cricketers is a restaurant serving Chinese food near All Bar One. It is family-friendly, has cheap price range and average customer rating.

WebNLG

input: *Allen Forrest | background | solo singer ► Allen Forrest | genre | pop music ► Allen Forrest | birthplace | Dothan, Alabama*

templates: *Allen Forrest is a solo singer. Allen Forrest performs Pop music. Allen Forrest was born in Dothan, Alabama.*

output: Allen Forrest is a solo singer who performs Pop music. He was born in Dothan, Alabama.

input: *Juan Perón | party | Labour Party (Argentina) ► Alberto Teisaire | inOfficeWhilePresident | Juan Perón ► Alberto Teisaire | nationality | Argentina ► Argentina | language | Spanish language*

templates: *Juan Perón belongs to the Labour Party Argentina. Alberto Teisaire was in office while Juan Perón was a president. Alberto Teisaire is from Argentina. Spanish language is spoken in Argentina.*

output: Alberto Teisaire is from Argentina, who was in office while Juan Perón was a president. He belongs to the Labour Party Argentina. Spanish language is spoken in Argentina.

disfluent

bad pronoun coreference

input: *Alfa Romeo 164 | relatedMeanOfTransportation | Fiat Croma ► Alfa Romeo 164 | assembly | Italy ► Italy | capital | Rome*

templates: *Alfa Romeo 164 is related to Fiat Croma. Alfa Romeo 164 was assembled in Italy. Italy's capital is Rome.*

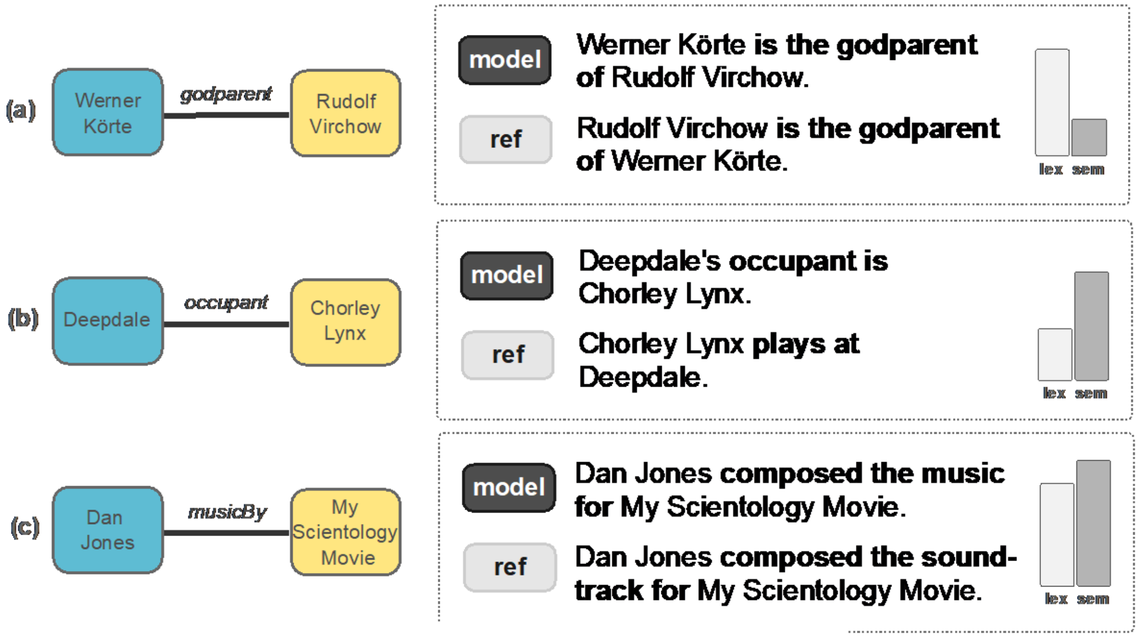
output: Alfa Romeo 164 was assembled in Italy's capital, Rome. It is related to Fiat Croma.

mixing unrelated facts

Describing relations with PLMs

(Kasner et al., 2023)
https://arxiv.org/abs/2210.07373

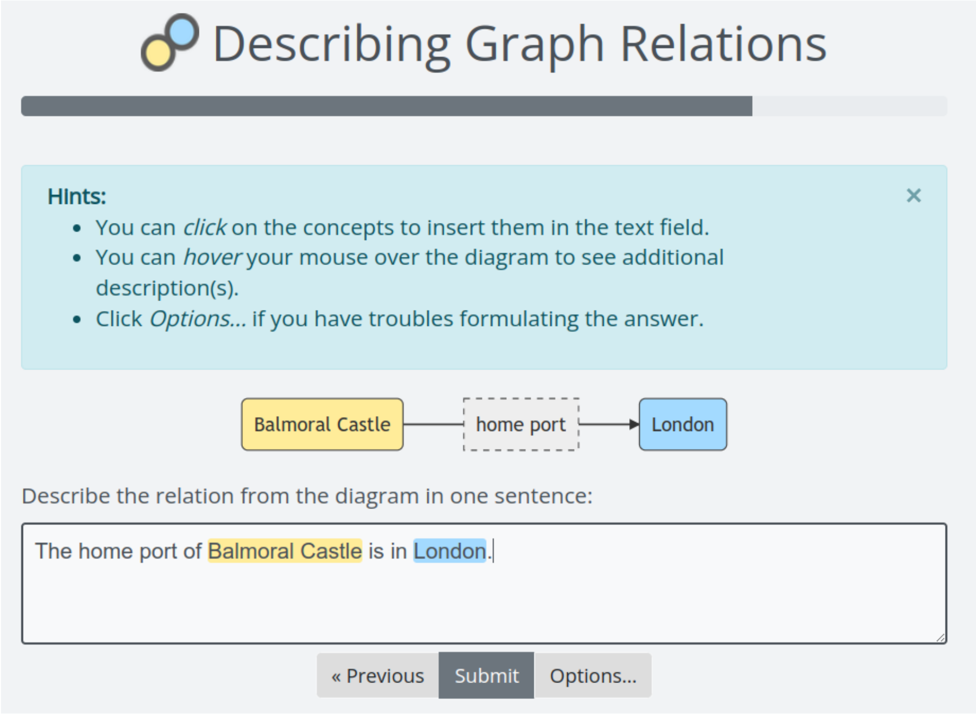
- Removing the **data → template step** in the pipeline
  - i.e. PLM to verbalize single triples
  - go **100% neural, zero-shot**
- Relations are most important
  - entities can be copied verbatim
- Relation labels often difficult
  - relation direction unclear
  - other label ambiguities
  - dependence on entities
- **How good are PLMs at this?**



relation	possible verbalization
is part of	X is part of Y.
duration	X lasted for Y.
platform	X is available on Y.
	X runs on Y.
country	X was born in Y.
	X is located in Y.
parent	X is the parent of Y.
	Y is the parent of X.
ChEMBL	X has an id Y in the ChEMBL database.

Rel2Text dataset

- Current data-to-text datasets unsuitable to test this
  - low number of distinct relations
  - few unseen in training set
- **New Rel2Text dataset**: 1.5k unique relations
  - source: Wikidata, YAGO, DBPedia
  - no train-test overlap
- **Crowdsourced** collection
  - 1-5 instances per relation
  - workers asked to rewrite relation as sentence
    - given relation labels & descriptions
  - **manual checks** for noise
    - 7.3k instances collected → 4k “clean”
  - ~ **hard** even for (untrained) people



Evaluating PLMs on Rel2Text

- Evaluation on unseen relations only
- Same PLM (BART), finetuned on different data
  - WebNLG = less diversity, more data
  - Rel2Text = many relations
  - Rel2Text with relation descriptions
  - Rel2Text with masked relation labels
    - guessing from entities only
- Finetuning works
  - Full Rel2Text best
  - Relation descriptions don't help much
  - WebNLG also OK (esp. on correctness)

	~overlap with human	~correctness	~fluency
Rel2Text	**BLEU**	**% Log. Entail**	**PPL↓ (GPT2)**
Human	-	-	5.88
Copy baseline	29.04	91.21	7.55
BART-WebNLG	41.99	89.39	5.65
BART-Rel2Text	52.54	91.85	5.89
+rel. descriptions	53.07	91.88	5.92
- rel. labels (guess)	42.53	57.26	5.66

Error Analysis

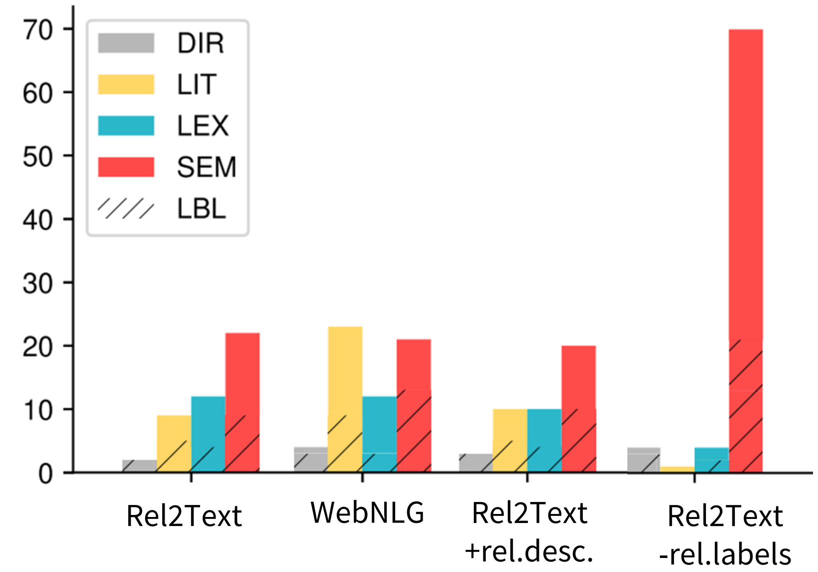
- 100 examples, multiple error classes

model errors

SEM	**semantic error**	Yousra Matine | *sport country* | Morocco	✗ Yousra Matine was born in Morocco. ✓ Yousra Matine plays for Morocco.
DIR	**swapped direction**	Kentucky Channel | *former broadcast network* | KET ED	✗ KET ED was broadcast on Kentucky Channel ED. ✓ The Kentucky Channel was broadcast on KET ED.
LIT	**verbalization too literal**	Vietnam Television | *first air date* | 1970-09-07	✗ The first air date of Vietnam Television was 1970-09-07. ✓ Vietnam Television first aired on 1970-09-07.
LEX	**lexical/grammar error**	RPG-43 | *used in war* | The Troubles	✗ RPG-43 was used in the The Troubles. ✓ The RPG-43 was used in the Troubles.
LBL	**label unclear**	General Motors Epsilon platform | *vehicle* | Cadillac XTS	✗ General Motors Epsilon is a vehicle similar to the Cadillac XTS. ✓ General Motors Epsilon platform is used in the Cadillac XTS.

data errors

- Near constant % of unclear labels
  - leading to SEM errors
- Still some "unprovoked" SEM errors
  - masked labels: much more
- Rel2Text → less LIT errors than WebNLG



Final Remarks

- **Rel2Text with PLMs viable**
  - comparable to templates in full pipeline
- Prompting LLMs ~ similar performance
  - GPT3 “templates” by Xiang et al.
- **Clear relation labels** are essential
  - even humans confused without them
  - additional descriptions help
- Ambiguities in data should be fixed prior to generation
- **Still >0% hallucinations** – semantics + alignments needed
  - work in progress

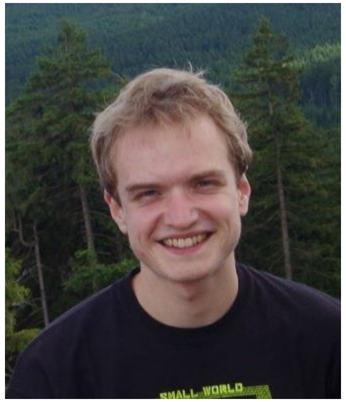
WebNLG	BLEU	Omission/ #facts	Hallucination/ #examples
Templates	37.18	0.000	0.000
Templates + 3-stage	42.92	0.051	0.148
BART/Rel2Text + 3-stage	44.63	0.058	0.166
GPT3 + 1-stage (Xiang et al.)	43.33	-	-

(Xiang et al., 2022) http://arxiv.org/abs/2210.04325

Thanks

Contact us:

Ondřej Dušek
odusek@ufal.mff.cuni.cz
https://tuetschek.github.io
@tuetschek



Zdeněk Kasner
kasner@ufal.mff.cuni.cz
http://ufal.cz/zdenek-kasner
@ZdenekKasner

Ioannis Konstas
i.konstas@hw.ac.uk
http://www.ikonstas.net/
@sinantie

References:

- Base pretrained LMs: (Kasner & Dušek, INLG/WebNLG 2020) https://aclanthology.org/2020.webnlg-1.20/
- Zero-shot pipeline: (Kasner & Dušek, ACL 2022) https://aclanthology.org/2022.acl-long.271/
- Rel2Text: (Kasner, Konstas & Dušek, EACL 2023) https://arxiv.org/abs/2210.07373

Supported by: European Research Council (ERC StG No. 101039303 NG-NLG), Charles University projects PRIMUS/19/SCI/10, GAUK 140320 and SVV 260575, Apple NLU Research Grant for Heriot-Watt University and Charles University.
Using LINDAT/CLARIAH-CZ Research Infrastructure resources (Czech Ministry of Education, Youth and Sports project No. LM2018101).

European Research Council
Established by the European Commission

Evaluating Data-to-text NLG

- **n-gram metrics** (BLEU, METEOR)
  - derived from MT, no good for accuracy
  - dubious even as measures for overall quality
- **Neural metrics** (BERTScore, BLEURT) mix accuracy & fluency
  - slightly better than n-gram, but still not ideal
- **SER** evaluation uses regex or exact match
  - tedious to make / inaccurate
  - does not translate to other datasets
- Proper evaluation means full NLU
  - pretrained LMs are good at NLU-like tasks → use them?

Checking for Errors in NLG Output: Natural Language Inference

- **NLI**: relation of premise (= starting point) & hypothesis (= relating text)
  - **E**ntailment = all hypothesis facts are included in premise
  - **N**eutral = not all hypothesis facts included, but no directly opposing facts
  - **C**ontradiction = premise is opposed by hypothesis

P: *Blue Spice is a pub in the riverside area.*

H$_1$: *Blue Spice is located in the riverside.* → **E**
H$_2$: *You can bring your kids to Blue Spice .* → **N**
H$_3$: *Blue Spice is a coffee shop.* → **C**

- We'll use a vanilla model trained for NLI
- Check entailment in both directions
  - data entails text = no hallucination + text entails data = no omission
- Use templates to represent data (same as previously)

(Dušek & Kasner, 2020)
https://www.aclweb.org/anthology/2020.inlg-1.19

Eval1: NLI Classification

(Dušek & Kasner, 2020)
https://www.aclweb.org/anthology/2020.inlg-1.19

1) Check for omissions

- premise = whole generated text
- hypothesis = each single fact, loop
  - → also checks which fact is omitted

2) Check for hallucination

- premise = concatenated facts
- hypothesis = whole generated text
  - can't easily split into simpler checks

- output:
  - 4-way – *OK, omission, hallucination, o+h*
  - 2-way – *OK, not_OK*
  - OK confidence (min. **E** confidence)
  - list of omitted facts

Blue Spice | eat_type | pub
Blue Spice | area | riverside

↓ NLG ↓

You can bring your kids to Blue Spice in the riverside area.

P: *You can bring your kids to Blue Spice in the riverside area.*

H_1: Blue Spice is a pub. C: 0.01 N: **0.97** E: 0.02

→ *omission*

H_2: Blue Spice is located in the riverside. C: 0.00 N: 0.01 E: **0.99**

→ *OK*

P: *Blue Spice is a pub. Blue Spice is located in the riverside.*

H: *You can bring your kids to Blue Spice in the riverside area.*

C: 0.00 N: **0.99** E: 0.01

→ *hallucination*

---

omission+hallucination

OK: 0.01 omitted: Blue Spice | eat_type | pub

Error Checking with NLI

- WebNLG & E2E data
  - comparison vs. human ratings (WebNLG) & SER regex script (E2E)
  - both datasets: default & backoff-only versions of templates

system	**WebNLG** data	**E2E data**	
		4-way	2-way
Accuracy / agreement	77.5%	91.1%	93.3%

- manual analysis: ca. ½ “errors” are in fact correct
  - annotation noise / SER script errors
  - noisy templates
  - edge cases (*high restaurant*)
  - stuff SER script doesn’t catch (*with full service*)