Data-to-text Generation with Neural Language Models

Ondřej Dušek

SCIA 2023
20.4.2023

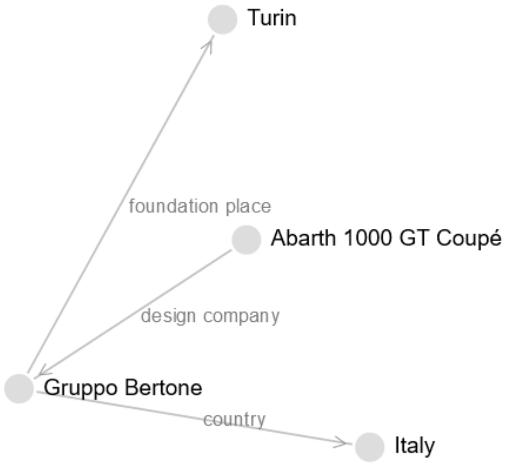
Thanks: Zdeněk Kasner, Ioannis Konstas

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

European Research Council
Established by the European Commission

Data-to-text NLG

- **data-to-text NLG** = verbalizing structured outputs
  - e.g. 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	
…				

→

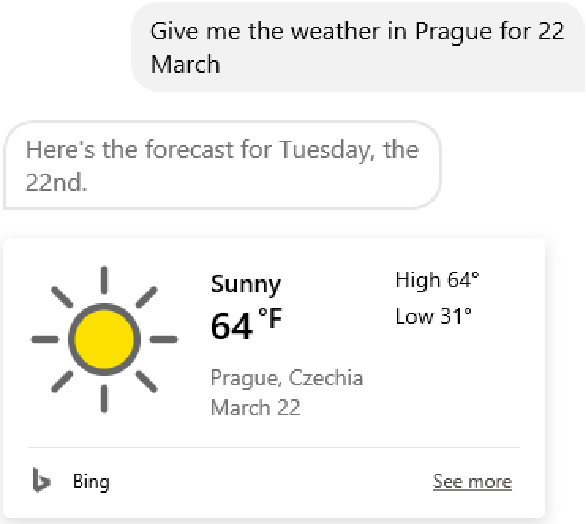
- *The Toronto Raptors, which were leading at halftime by 10 points (54-44), defeated the Dallas Mavericks by 8 points (94-86).*

…

- *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

NLG Approaches

- **hand-written prompts** ("canned text")
  - trivial – hard-coded, doesn't scale (good for IVR/DTMF phone systems)
- **templates** ("fill in blanks")
  - simple, but much more expressive
  - can scale if done right, still laborious
  - most commercial systems today!

[name] *is a* **[eat_type]** *in the* **[area]** *area.*

name = Blue Spice
eat_type = pub
area = riverside

+

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

- **grammars & rules**
  - experimental, pipelines, more expressive but more laborious
- **machine learning** (neural LMs → →)

Neural NLG

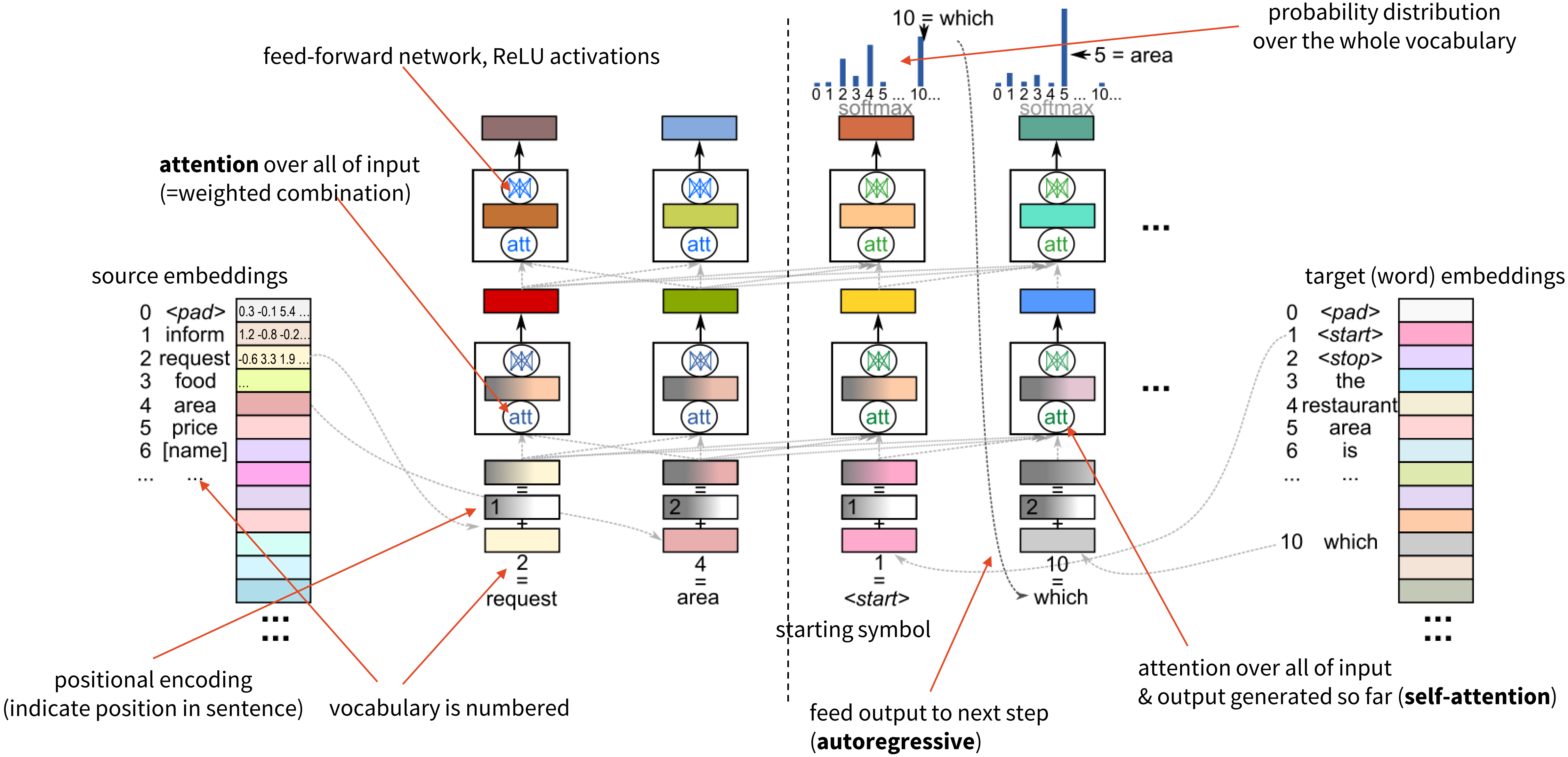
- 1 step, **end-to-end**
  - feed input data (linearized)
  - generates text **autoregressively**: word-by-word, left-to-right
- **Transformer** neural architecture
  - **encoder** (takes input) – **decoder** (produces output)
  - alt.: decoder-only (both input & output)
- **Trained** fully from input-output pairs
  - Needs a lot of training data (~10k range, 10x more than before)
- Much more **fluent** outputs
- Opaque & has **no guarantees on accuracy**
  - used essentially as a black box, internals unknown

Neural NLG: Transformer Models

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

1) encoder: encode linearized data

2) decoder: decode text word-by-word



Neural NLG: Training

- Trained to produce sentences from data
  - replicate exact word at each position (given gold context)
- **Supervised** learning
  - initialize model with random parameters
  - **classification**: 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

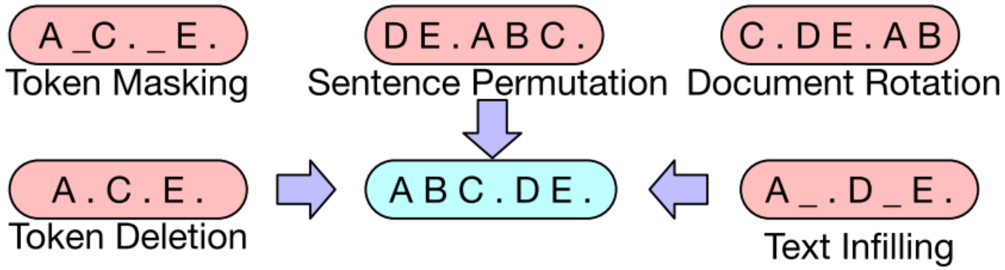
reference:

Blue Spice is expensive

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

Neural NLG: Pretraining + Finetuning

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



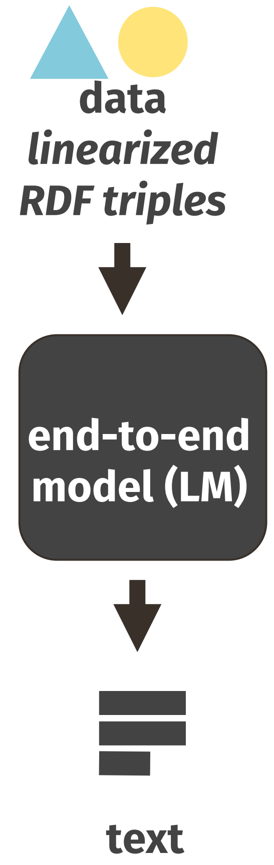
(Lewis et al., 2020)
https://www.aclweb.org/anthology/2020.acl-main.703

- Models free for download (https://huggingface.co/)
  - BERT/RoBERTa, GPT-2, BART, T5…
  - 100k-1B parameters – runs easily on regular GPUs

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 LM
  - direct data-text mapping: black box
  - needs domain-specific data
    - scarce (~10k max)
    - noisy (crowdsourced)
  - no guarantees on accuracy

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



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

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)

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

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!

Large language models (LLMs): Pretrain & prompt

- 10-100B parameters
  - harder to run in-house (OPT, BLOOM, LlaMa) or not free (GPT-3, ChatGPT, LaMDa)
- architecture mostly the same
- **prompting:** context / examples / question → reply
  - typically no need to finetune
  - finetuning: expensive, less effect
- hard to control

GPT3.5 data-to-text

Write a short description based on data.

Example: Input data: name = Cocum, eatType = coffee shop, food = English, priceRange = high, customer rating = average, familyFriendly = yes
Output description: An English coffee shop named Cocum, though child-friendly, has only an average rating given its high prices.

Input data: eatType = pub, food = Indian, priceRange = cheap, name = Cotton, near = Portland Arms

Output description: Cotton is a cheap Indian pub located near the Portland Arms.

https://platform.openai.com/playground/

https://jlibovicky.github.io/2023/02/20/Question-and-Answers-about-ChatGPT-and-LLMs.html

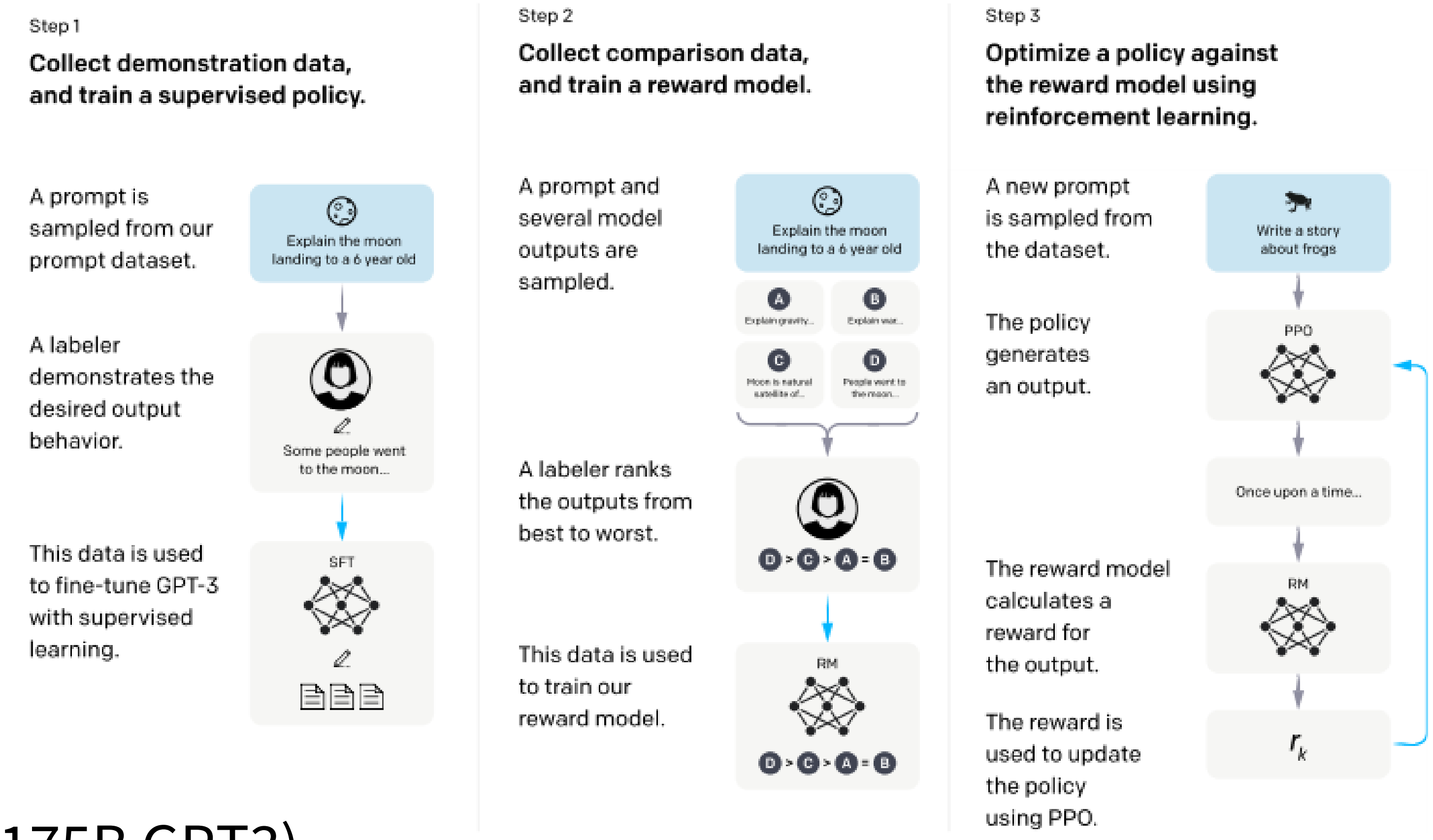
Instruction Tuning & RLHF

(Ouyang et al., 2022) http://arxiv.org/abs/2203.02155
https://openai.com/blog/chatgpt/ , https://www.youtube.com/watch?v=0A8ljAkdFtg
https://crfm.stanford.edu/2023/03/13/alpaca.html , https://open-assistant.io/

- InstructGPT/ChatGPT, Alpaca, OpenAssistant
- Multi-step training (start from a LLM)

1) finetuning
   - standard, but better data ("in-domain")
2) evaluation/ranker model
   - lots of rankings
   - used to train RL reward model
3) RL with reward model

- better replies,
  can be made safer
- more efficient (1.3B InstructGPT > 175B GPT3)



ChatGPT Example

- Impressive text editing & composition skills
- Hallucinates a lot (only using input / memorized info)
  - trained to be convincing, not truthful

Who is Ondřej Dušek?

14, depending how you count

Ondřej Dušek is a researcher and academic in the field of natural language processing (NLP). He is currently an Assistant Professor at the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

Dušek's research focuses on developing computational models and algorithms for natural language generation, dialogue systems, and machine learning approaches for NLP tasks. He has published several research papers in top-tier NLP conferences and journals, and he has also served as a reviewer and program committee member for these conferences.

no other tasks than generation & dialogue

In addition to his research, Dušek is also actively involved in teaching and mentoring students in the field of NLP. He has supervised several Master's and PhD theses, and he teaches courses on topics such as computational linguistics, NLP, and machine learning.

no PhD graduate yet

I only teach dialogue systems

Compose a limerick on how to exit Vim.

There once was a programmer in Vim
Whose exit strategy was looking grim
They tried to escape
But couldn't find the right shape
Until they remembered to type ":q!" and win!

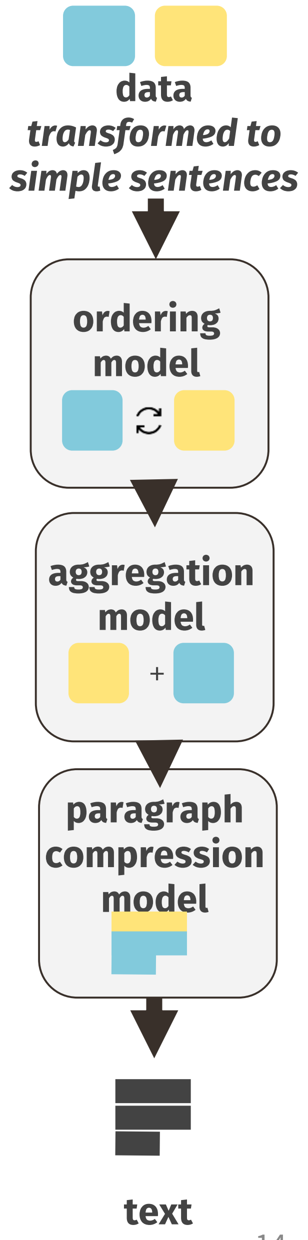
Nice, now try it while speaking like a pirate

Arrr, listen ye mateys to me tale,
Of a sailor who be stuck in Vim's jail,
He couldn't leave the text editor,
No matter how he did endeavour,
But then he remembered ":q!" and set sail!

Text-to-Text: LMs for Editing Only

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

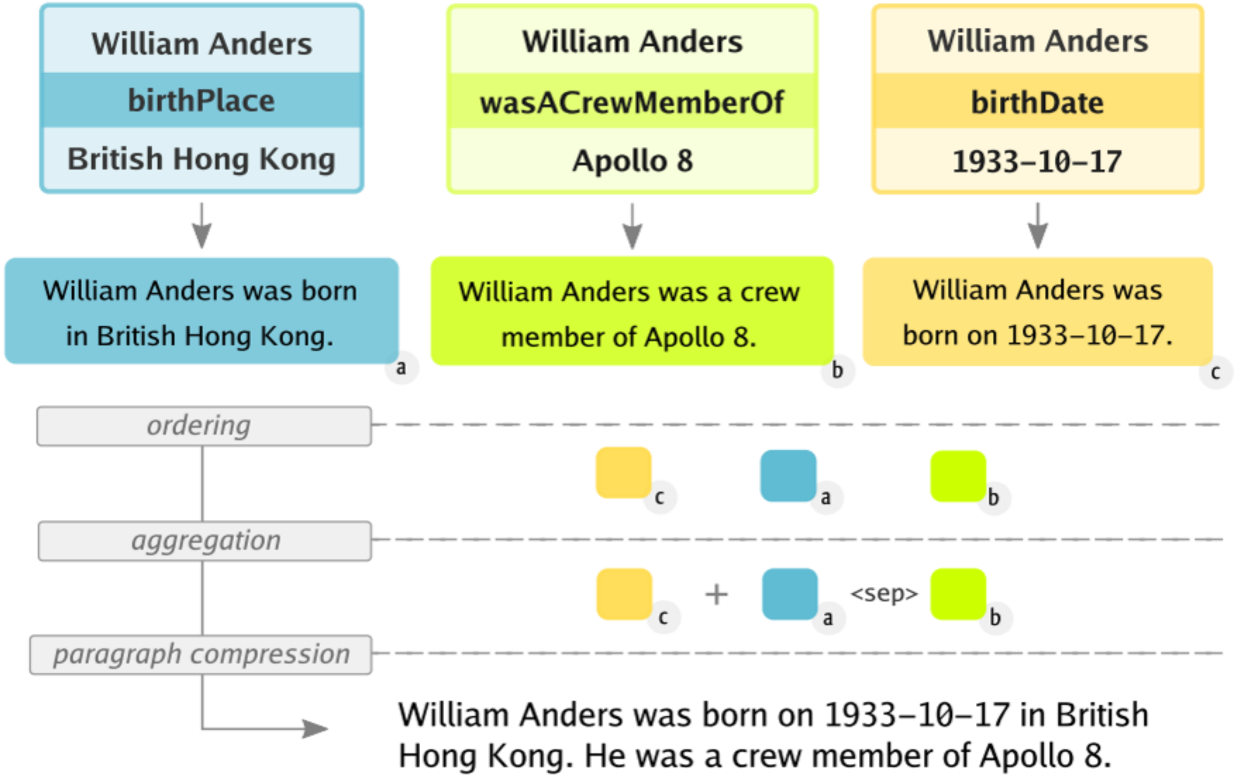
- Represent input data by templates
  - handcrafted, but not so many needed (1 per input relation)
  - entities inserted verbatim, don't need to be fluent
- 3 neural LMs to **fuse & rephrase:**
  - All text-to-text steps (=editing only, making text more fluent)
  - **1) order** (put related stuff together)
  - **2) aggregate** (into sentences)
  - **3) compress** (produce shorter sentences)
- Less space for semantic errors
  - Use LMs for what they're pretrained for & good at – fluency
- Can use large general-domain data
- Works **zero-shot** – needs no in-domain data (just the templates)



Text-to-Text LM Editing Pipeline

1) Templates
2) **Ordering**
   - BART LM with a pointer network
3) **Aggregation**
   - RoBERTa LM + classif. same/new sent.
4) **Paragraph compression**
   - BART LM – generation

- **WikiFluent:** synthetic training data
  - 1M instances, domain-general (Wikipedia)
  - human-written targets
  - synthetic sources resembling templates



Results

- Better than previous neural or 1-step
- Still quite fluent
- Still not 100% correct

~overlap with human

~correctness

WebNLG data	BLEU	Omission/ #facts	Hallucination/ #examples
Older neural systems	45.13	0.237	0.202
Templates	37.18	0.000	0.000
Templates + 1-step	39.08	0.071	0.204
Templates + 3-step	42.92	0.051	0.148

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 LMs

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

- Removing the **data → template step** in the pipeline
  - i.e. LM to verbalize single triples
  - go **100% neural, zero-shot**
- Text-to-text easier than data-to-text
  - expressing relations difficult
- **How good are LMs at this?**
- **Rel2Text**: dataset to test this
  - current sets are not diverse enough
  - 1.5k relations / 4k examples from Wikidata/YAGO/DBPedia
  - crowdsourced + manual checks
- It's hard for people too
  - our checks removed ~45% data

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.

Evaluating LMs on Rel2Text

- On unseen relations only
- **Finetuning** BART
  - Rel2Text works well
  - WebNLG also OK (esp. on correctness)
- **Prompting** ChatGPT
  - requires carefully crafted prompts
  - chattier outputs (~less control)
- Error analysis
  - **Unclear relation labels lead to semantic errors**
  - Still some “unprovoked” semantic errors
  - BART + Rel2Text & ChatGPT produce nicer, less literal verbalizations

	~overlap with human	~correctness	~fluency
Rel2Text data	**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
ChatGPT	38.23	88.58	5.68

Final Remarks

- **Editing Pipeline > End-to-end**
  - Can be fully neural
  - BART/Rel2Text as good as templates
  - Prompting LLMs ~ similar performance
    - GPT3 “templates” by Xiang et al.
- **Clear inputs** are essential
  - even humans confused without them
  - often need more detail
- **Still >0% hallucinations** for any method
  - detailed semantics + alignments needed
  - work in progress

~overlap with human

~correctness

WebNLG	BLEU	Omission/ #facts	Hallucination/ #examples
Older neural systems	45.13	0.237	0.202
Templates	37.18	0.000	0.000
Templates + 3-step	42.92	0.051	0.148
BART/Rel2Text + 3-step	44.63	0.058	0.166
GPT3 + 1-step (Xiang et al.)	43.33	-	-

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

Thanks

Contact me:

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

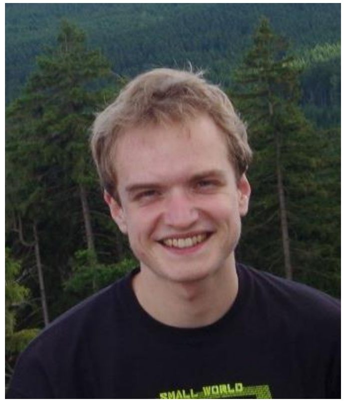
These slides:

http://bit.ly/scia2023-od

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

Thanks:



Zdeněk Kasner
@ZdenekKasner

Ioannis Konstas
@sinantie

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