Data-to-text generation can be formulated as a sequence of trainable **text-to-text operations**

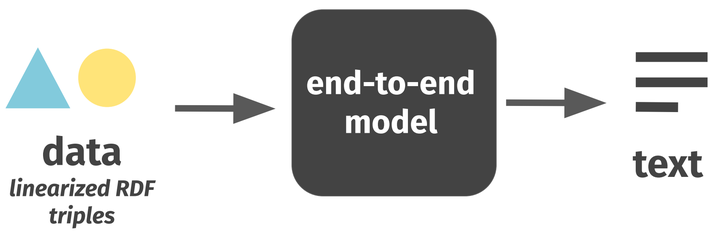
Neural Pipeline for Zero-Shot Data-to-Text Generation

Zdeněk Kasner *kasner@ufal.mff.cuni.cz*
Ondřej Dušek *odusek@ufal.mff.cuni.cz*

CHARLES UNIVERSITY
ÚFAL

Naive approach

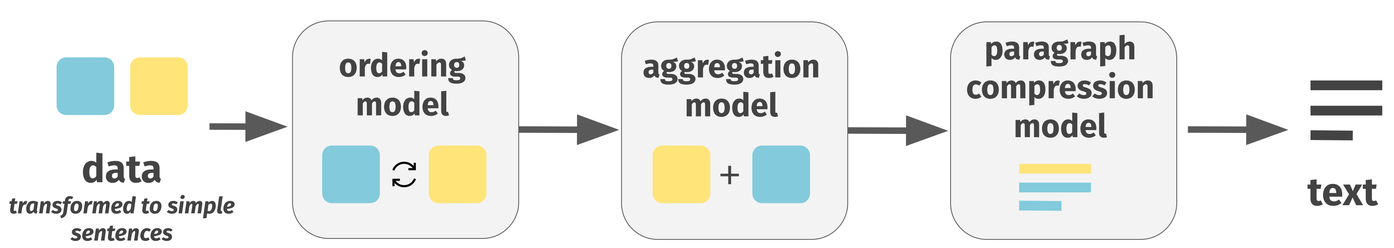
x



- **x** low number of training examples (~10k)
- **x** direct mapping from data to text → black box
- **x** noise from crowdsourcing → omissions, hallucinations
- **x** needs domain-specific data

Our approach

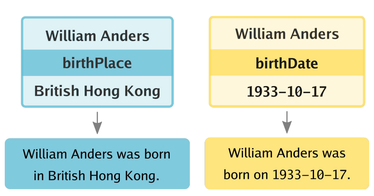
✓



- **+** large-scale training data for each model (~1M)
- **+** interpretable intermediate operations over natural language
- **+** preserving the input semantics → no incentives for omissions or hallucinations
- **+** general-domain training data automatically created from Wikipedia

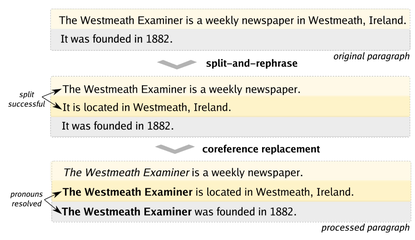
Ingredients

Transforming data to sentences



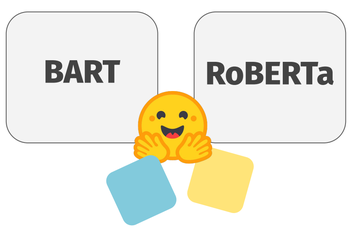
- simple handcrafted templates for RDF predicates
- balances manual workload and controllability

WikiFluent corpus



- 934k sets of simple sentences synthesized from first paragraphs on English Wikipedia
- split-and-rephrase model + coreference resolution
- used to train all pipeline modules
- filtered version: semantic accuracy checked using RoBERTa-MNLI

Pipeline modules



- **ordering (ord):** BART with pointer network *(Calizzano et al., 2021)*
- **aggregation (agg)**: RoBERTa-large *(Liu et al., 2019)* with token classification head
- **paragraph compression (PC)**: BART-base *(Lewis et al., 2020)*

Experiments

Zero-shot data-to-text generation

1. training the ordering, aggregation and PC modules on our **WikiFluent** corpus
2. handcrafting templates for 354 predicates in **WebNLG** *(Gardent et al., 2017)* and 8 predicates in **Cleaned E2E** *(Dušek et al., 2019)* datasets
3. applying the pipeline on the datasets in **zero-shot** mode (no finetuning)
   + ablation study: pipeline without *agg* (2-stage) or *ord+agg* (1-stage) modules

Results

WebNLG			
system	**BLEU**	**O**	**H**
Baseline (concat templates)	37.18	0.000	0.000
Melbourne - WebNLG Challenge 2017	45.13	0.237	0.202
SOTA (Ke et al., 2021)	66.14	-	-
Ours (2-stage, filt.)	43.49	0.146	0.096

E2E			
system	**BLEU**	**O**	**H**
Baseline (concat templates)	24.19	0.000	0.000
TGen (Dušek and Jurčíček, 2015)	40.73	0.016	0.083
SOTA (Harkous et al., 2020)	43.60	-	-
Ours (3-stage, full)	36.04	0.001	0.001

O = omissions / # triples, H = hallucinations / # examples, computed using RoBERTa-MNLI *(Dušek and Kasner, 2020)*

Output analysis

- output texts generally adhere to the initial templates - similar wording, facts are not deleted / added → prevents omission & hallucination
- improved fluency compared to templates only (better ordering, sentence fusion, using coreferences, minor rephrasing)
- **problems:**
  - over-eager sentence fusion may change the semantics
  - creating the templates manually limits the applicability to larger datasets

Example

input: (Allen Forrest; *background*; solo singer), (Allen Forrest; *genre*; Pop music), (Allen Forrest; *birthPlace*; Dothan, Alabama)
facts: Allen Forrest is a solo singer. Allen Forrest performs Pop music. Allen Forrest was born in Dothan,Alabama.
model: Allen Forrest is a solo singer who performs Pop music. He was born in Dothan, Alabama.
reference: *Born in Dothan, Alabama, Allen Forrest has a background as a solo singer and was a pop artist.*

Presented at ACL 2022, Dublin.

https://github.com/kasnerz/zeroshot-d2t-pipeline

Supported by Charles University projects GAUK 140320, SVV 260575 and PRIMUS/19/SCI/10, an Apple NLU Research Grant for Heriot-Watt University and Charles University, and by the European Research Council (No. 101039303 NG-NLG).