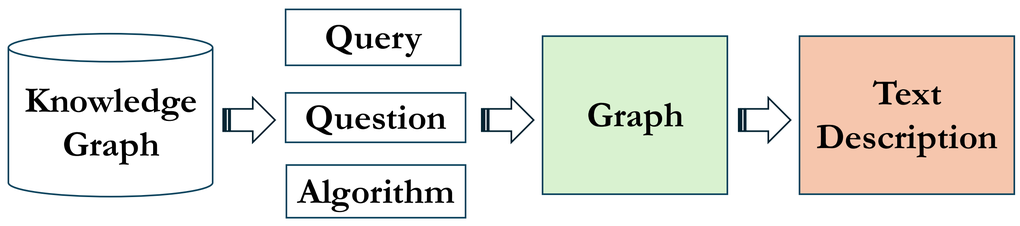
Hallucination Mitigation in Natural Language Generation from Large-Scale Open-Domain Knowledge Graphs

Xiao Shi, Zhengyuan Zhu, Zeyu Zhang, Chengkai Li

iDiR Department of Computer Science and Engineering
The University of Texas at Arlington

https://idir.uta.edu/

Need for Graph-to-Text



- Operations on knowledge graph (e.g., query, question-answering, data mining) lead to graph fragments which can be challenging for end users to comprehend.
- Natural language narration can help tackle this usability challenge.
- Fine-tuning PLMs produced STOA results on WebNLG (Ribeiro et al., 2021; Wang et al., 2021; Clive et al., 2021) and DART (Aghajanyan et al., 2021)

Hallucination in Graph-to-Text

- Output texts may contain fabricated facts not present in input graphs.

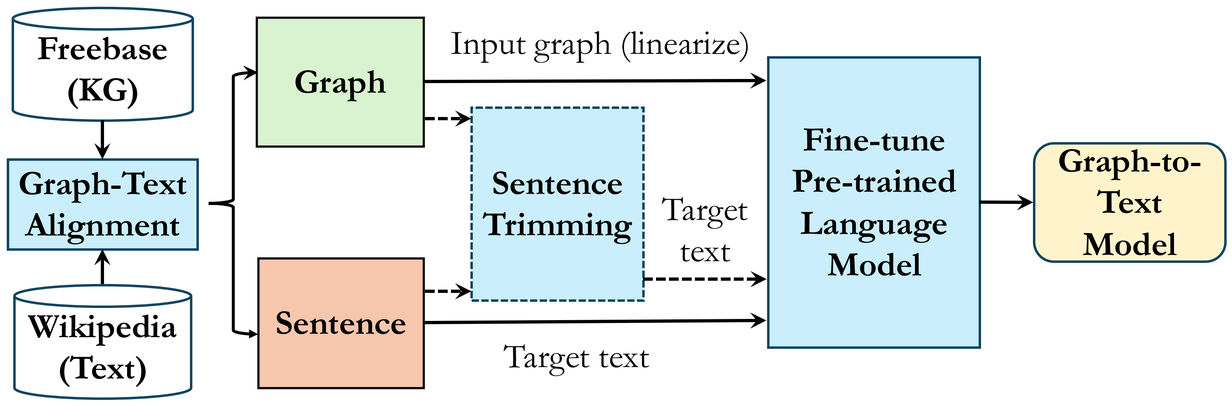
Goldie Gets Along is a 1951 American comedy film directed by Malcolm St. Clair (filmmaker) and starring Lili Damita and Charles Morton (actor)

Output / Input: **Graph-to-Text Model (Finetuned T5/BART)**

[S] Goldie Gets Along [P] film directed by [O] Malcolm St. Clair (filmmaker) [S] Goldie Gets Along [P] film actor [O] Lili Damita [S] Goldie Gets Along [P] film actor [O] Charles Morton (actor)

Linearize: Malcolm St. Clair (filmmaker) —*directed by*— Goldie Gets Along —*actor*— Lili Damita; Goldie Gets Along —*actor*— Charles Morton (actor)

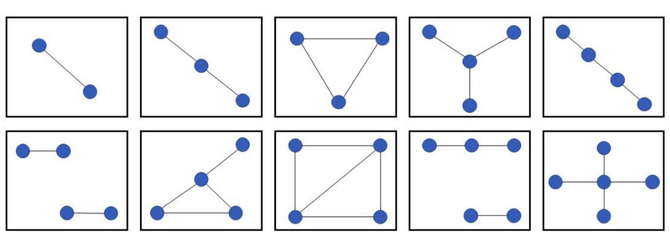
GraphNarrator: Large-Scale Graph-to-Text



GraphNarrative Dataset

GraphNarrative	Existing Graph-to-Text Datasets
Among the largest	Hard to scale human-written training graph-text pairs
Large linguistic variation	Hand-crafted text may follow monotonous templates
Diverse graph shapes	Largely limited to star graphs
Open-domain	Some focus on specific domains

Dataset	Text source	Domain	Instances	Entities	Triples	Relation	Star Graphs
WebNLG	human	15 DBpedia categories	25K	2K	3K	354	57%
DART	human	N/A	38K	27K	32K	3,834	83%
AGENDA	automatic	scientific research	40K	159K	177K	7	2%
EventNarrative	automatic	events	224K	305K	649K	672	94%
TEKGEN	automatic	open domain	7,895K	4,856K	11,373K	663	96%
GraphNarrative	automatic	open domain	8,769K	1,853K	15,472K	1,724	22%



10 most frequent graph shapes among 7,920 distinct shapes

Hallucination Mitigation: Sentence Trimming

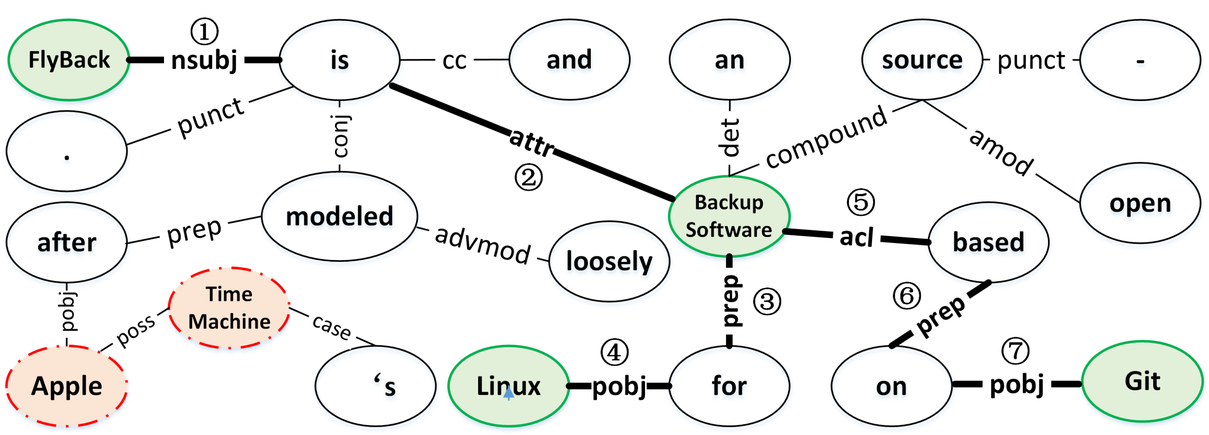
- Cause of hallucination: Inconsistency between graph and text in training pairs
- Idea for mitigating hallucination: Eliminating portions of the sentence not present in the graph while preserving the sentence's main idea

Graph

Backup Software —*software genre*— FlyBack —*basis*— Git; FlyBack —*operating system*— Linux

Sentence

FlyBack[1] is[2] an[3] open[4]-[5]source[6] **backup software**[7] for[8] **Linux**[9] based[10] on[11] **Git**[12] ~~and[13] modeled[14] loosely[15] after[16]~~ **~~Apple~~**[17]'s[18] **~~Time Machine~~**[19].[20]



Dependency parsing tree of the original sentence

Experiments and Results

Model	Sentence Trimming	GraphNarrative			TEKGEN		
		BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
BART-large	w/o	32.35	17.45	37.12	41.51	23.62	47.13
BART-large	w/	46.04	24.35	49.69	48.32	29.90	57.50
T5-large	w/o	22.22	17.16	36.78	43.03	24.21	48.05
T5-large	w/	45.12	24.77	50.44	49.83	30.52	58.25

Model performance on GraphNarrative and TEKGEN

- Fine-tuning T5-large model attained the best performance across most metrics.
- Models consistently perform better with sentence trimming than without.

Sentence Trimming	Hallucinated Entities	Missed Entities	Hallucinated Relations	Missed Relations	Grammar
w/o	1.163	0.003	1.340	0.040	4.793
w/	0.306	0.003	0.453	0.083	4.613

Human evaluation of GraphNarrative quality

Sentence Trimming	Hallucinated Entities	Missed Entities	Hallucinated Relations	Missed Relations	Grammar
w/o	1.643	0.063	1.363	0.240	4.613
w/	0.260	0.056	0.300	0.370	4.356

Human evaluation of generated sentences

- Sentence trimming effectively reduced hallucinated entities and relations.
- Only a slight decline in the grammar score.

Model	BLEU			METEOR			chrF++		
	all	seen	unseen	all	seen	unseen	all	seen	unseen
(Ribeiro et al., 2021)	59.70	64.71	53.67	44.18	45.85	**42.26**	75.40	78.29	72.25
(Wang et al., 2021)	60.56	66.07	53.87	44.00	46.00	42.00	-	-	-
GNST-T5 (ours)	**61.46**	**66.49**	**55.35**	**44.30**	**46.23**	42.08	**76.20**	**79.35**	**72.76**

Further fine-tuning results on WebNLG

- Fine-tuned T5-large model using GraphNarrative with sentence trimming achieved state-of-the-art results when further fine-tuned on WebNLG

Model	WebNLG			DART		
	BLEU	METEOR	chrF++	BLEU	METEOR	chrF++
T5-large	4.01	9.54	24.64	3.44	7.93	23.17
GN-T5	21.38	31.82	56.83	19.35	27.35	50.41
GNST-T5	27.6	32.27	56.81	19.42	28.07	50.96

Zero-shot learning results on WebNLG and DART

- GraphNarrative dataset can improve PLM's generalization ability

Contributions

- GraphNarrative: new dataset to fill the gap between existing datasets and large-scale real-world settings
- The first to quantify hallucinations in graph-to-text models
- Sentence trimming: novel approach for mitigating hallucination
- Comprehensive experiments and evaluations on GraphNarrative's quality and sentence trimming's effectiveness

https://github.com/idirlab/graphnarrator

GitHub

EMNLP 2023