UNIVERSITY OF TEXAS ARLINGTON

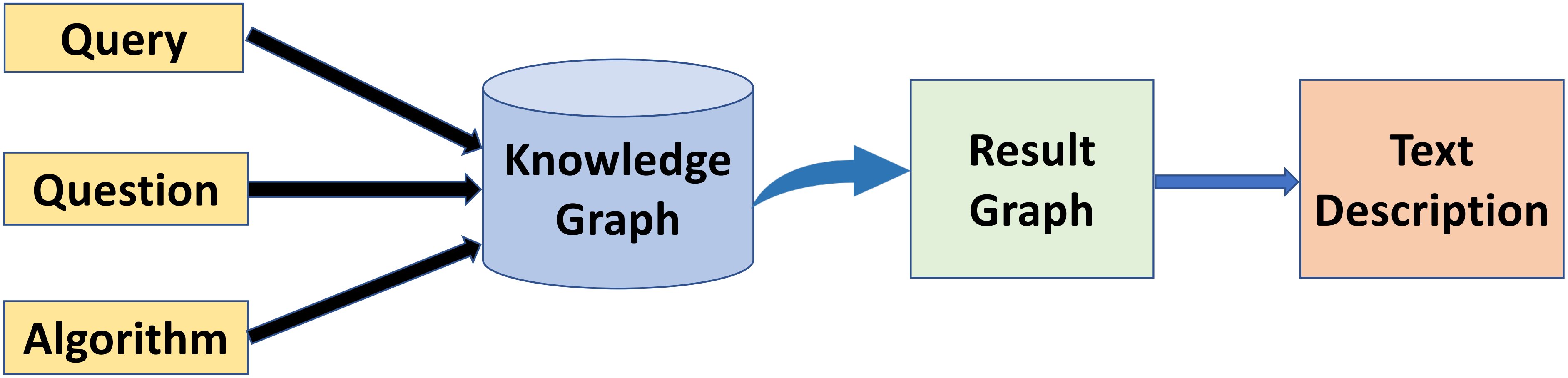
Hallucination Mitigation in Natural Language Generation from Large-Scale Open-Domain Knowledge Graphs

Xiao Shi, Zhengyuan Zhu, Zeyu Zhang, Chengkai Li

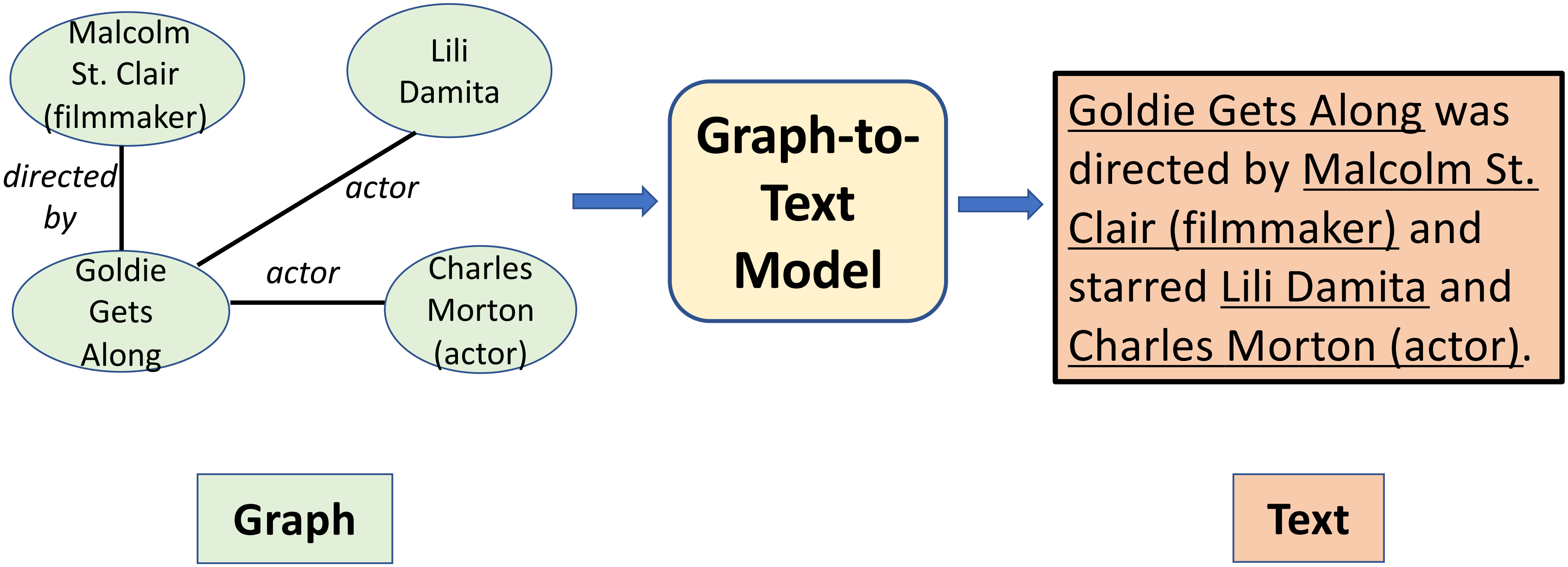
The Innovative Data Intelligence Research Laboratory (IDIR Lab)

EMNLP 2023

Motivation: Need for Narrating Graph Fragments



Graph-to-Text Model



Malcolm St. Clair (filmmaker)

Lili Damita

directed by

actor

actor

Goldie Gets Along

Charles Morton (actor)

Graph-to-Text Model

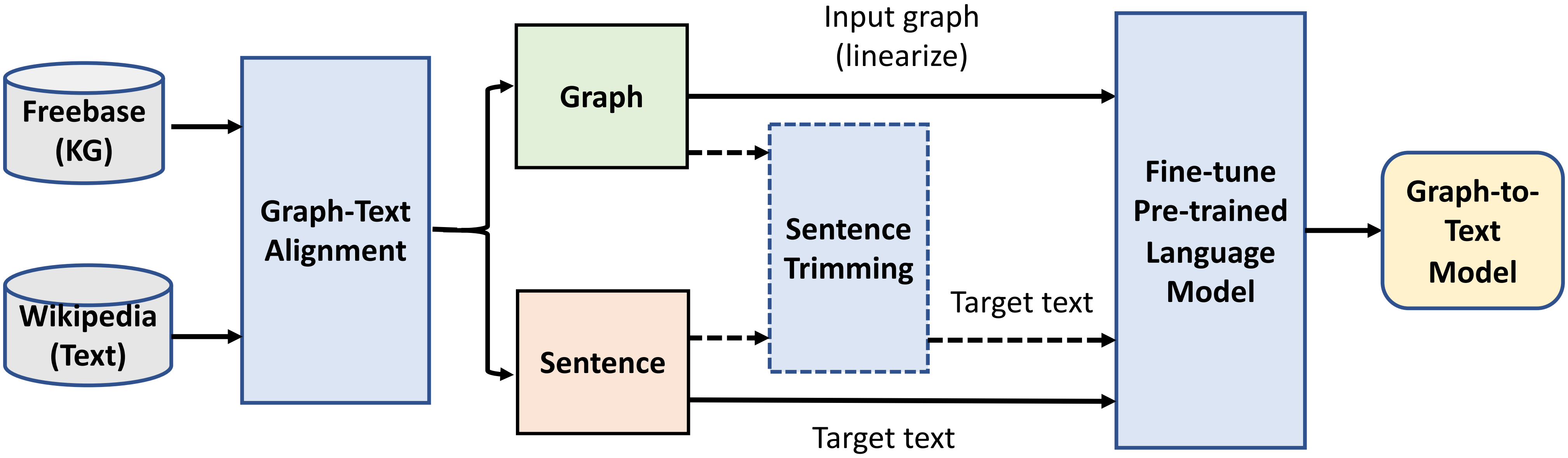
Goldie Gets Along was directed by Malcolm St. Clair (filmmaker) and starred Lili Damita and Charles Morton (actor).

Graph

Text

Overview of System

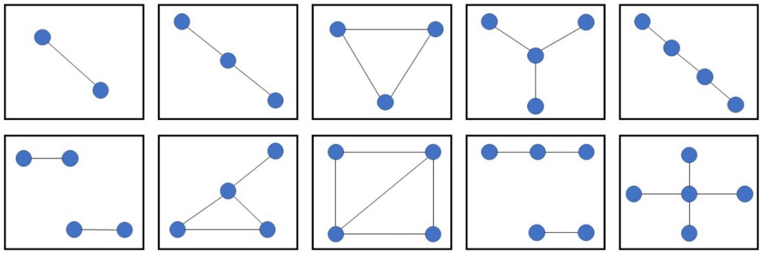
IDIR Department of Computer Science and Engineering
The University of Texas at Arlington



GraphNarrative vs. Existing Datasets

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

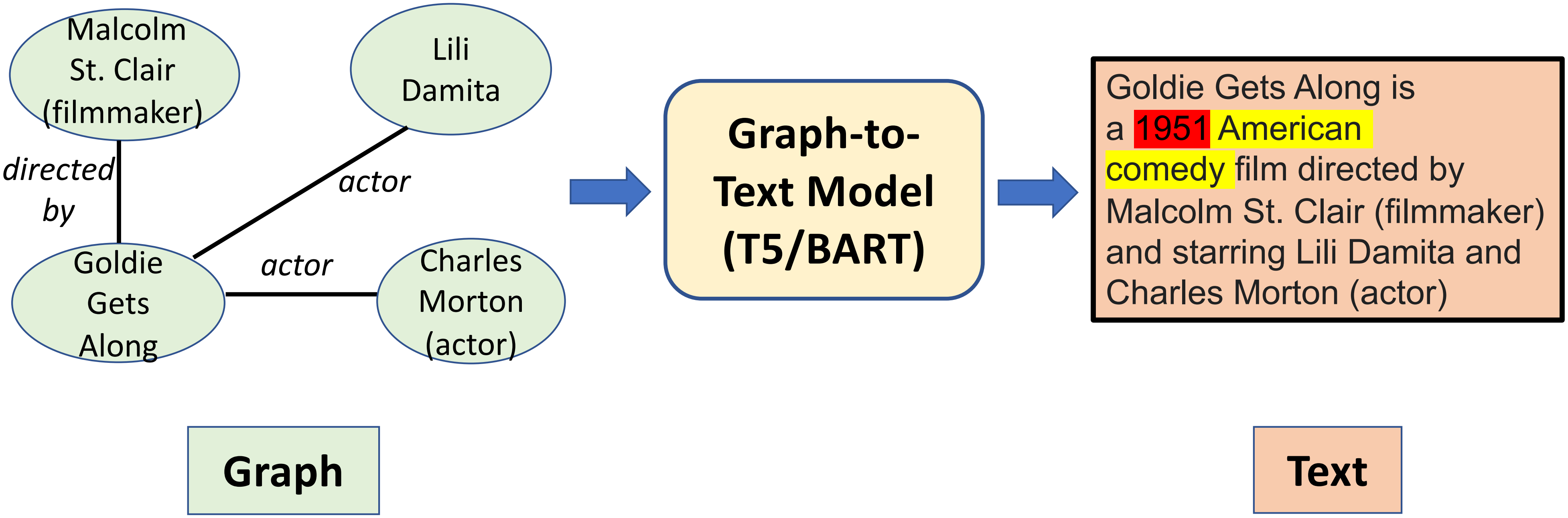
- ○ Among the largest
- ○ Diverse graph shapes
- ○ Large linguistic variation
- ○ Open-domain



10 most frequent graph shapes among 7,920 distinct shapes

Dataset	Text source	Domain	Instances	Entities	Triples	Relation	Star Graphs
WebNLG	human	15 DBpedia categories	25,298	2,730	3,221	354	57%
DART	human	N/A	38,391	27,000	32,139	3,834	83%
AGENDA	automatic	Scientific research	40,720	159,691	177,568	7	2%
EventNarrative	automatic	Events	224,428	305,685	649,337	672	94%
TEKGEN	automatic	Open domain	7,895,789	4,856,439	11,373,838	663	96%
GraphNarrative	automatic	Open domain	8,769,634	1,853,752	15,472,249	1,724	22%

Hallucination Problem



Malcolm St. Clair (filmmaker)

Lili Damita

Goldie Gets Along

Charles Morton (actor)

directed by

actor

actor

Graph-to-Text Model (T5/BART)

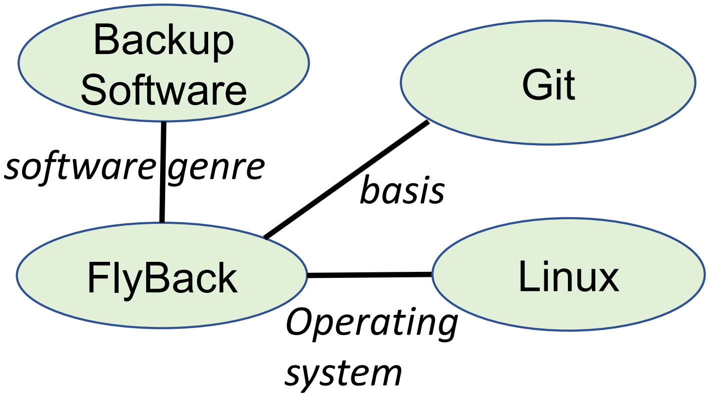
Goldie Gets Along is a 1951 American comedy film directed by Malcolm St. Clair (filmmaker) and starring Lili Damita and Charles Morton (actor)

Graph

Text

Cause of Hallucination

Inconsistency between graph and text in the training pairs



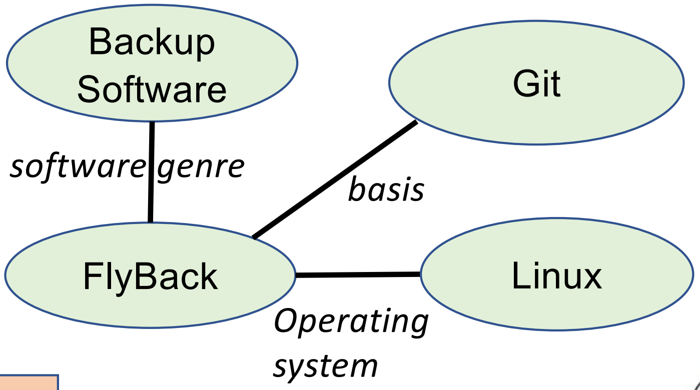
FlyBack is an open-source backup software for Linux based on Git and modeled loosely after Apple's Time Machine.

Graph

Text

Sentence Trimming

Graph

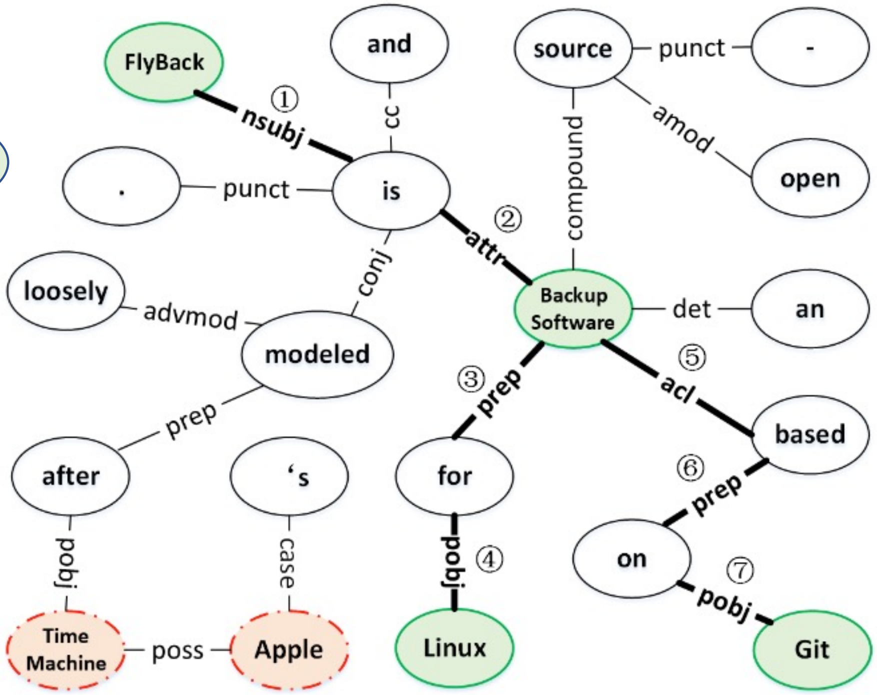


Original Text

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]

Trimmed Text

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 original sentence

Experiment Setup

Datasets

- GraphNarrative: our new dataset
- TEKGEN: most similar with ours, but mostly star graphs
- WebNLG: human annotated benchmark dataset
- DART: human annotated benchmark dataset

Models

- T5: small / base/ large - 60M / 220M / 770M parameters
- BART: base / large - 140M / 400M parameters

Evaluation Metrics

- **Automatic evaluation:** BLEU, METEOR, chrF++
- **Human evaluation**
  - **Inconsistency**

    # *hallucinated entities:* entities not in graph but in sentence

    # *missed entities:* entities in graph but not in sentence

    # *hallucinated relations:* relations not in the graph but in sentence

    # *missed relations:* relations in graph but not in sentence
  - **Grammar**

    5: no grammatical errors

    4: one grammatical error

    3: two to three grammatical errors

    2: four to five grammatical errors

    1: more than five errors

Model Performance

Model	Sentence Trimming	BLEU	METEOR	chrf++
BART-large	w/o	32.35	17.45	37.12
BART-large	w/	46.04	24.35	49.69
T5-large	w/o	22.22	17.16	36.78
T5-large	w/	45.12	24.77	50.44

Model Performance on GraphNarrative

Model	Sentence Trimming	BLEU	METEOR	chrF++
BART-large	w/o	41.51	23.62	47.13
BART-large	w/	48.32	29.90	57.50
T5-large	w/o	43.03	24.21	48.05
T5-large	w/	49.83	30.52	58.25

Model Performance on TEKGEN

○ Fine-tuning the T5-large model attained the best performance across most metrics.

○ Models consistently perform better with sentence trimming than without.

Human Evaluation of GraphNarrative Quality and Generated Sentences

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

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

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

Further Fine-Tuning Results on WebNLG

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**

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

Zero-shot Leaning Results on WebNLG and DART

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

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

GitHub https://github.com/idirlab/graphnarrator