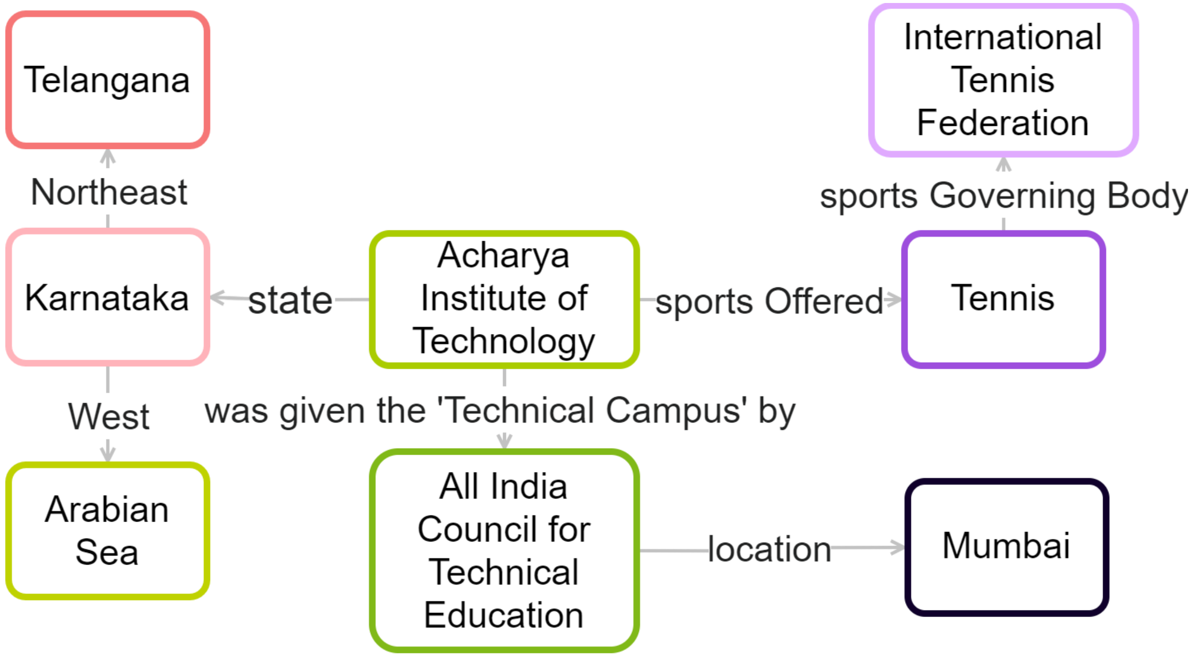
Stage-wise Fine-tuning for Graph-to-Text Generation

Qingyun Wang[1], Semih Yavuz[2], Victoria Lin[3], Heng Ji[1], Nazneen Rajani[2]

[1]University of Illinois at Urbana-Champaign, [2]Salesforce Research, [3]Facebook AI

The programs and data are publicly available at:
[https://github.com/ EagleW/Stage-wise-Fine-tuning](https://github.com/EagleW/Stage-wise-Fine-tuning)

Task



The Acharya Institute of Technology in Karnataka state was given Technical Campus status by All India Council for Technical Education in Mumbai. The school offers tennis which is governed by the International Tennis Federation. Karnataka has the Arabian Sea to its west and in the northeast is Telangana.

Motivation and Proposed Component

- How do humans describe structured data?
  - Factual knowledge from daily readings
  - Generalize linguistic expressions for fact description
  - Capture interdependence among facts
- Can machines repeat such a process?
  - Transformer-based language model pre-trained on large corpus
  - Fine-tune pre-trained language model on Wikipedia
  - Position embeddings to cover the structure of the graph

Token Embeddings	[CLS]	S|	Karnataka	P|	Northeast	...
Position Embeddings	POS_0	POS_1	POS_2	POS_3	POS_4	...
Triple Role Embeddings	ROL_0	ROL_1	ROL_1	ROL_1	ROL_2	...
Tree-level Embeddings	LV_0	LV_2	LV_2	LV_2	LV_2	...

Model Architecture: Positional Embeddings

- Given an RDF graph with multiple relations $G = \{(s_1, r_1, o_1), \ldots, (s_n, r_n, o_n)\}$, we flatten the graph as a concatenation of linearized triple sequences $|S\ s_1\ |P\ r_1\ |O\ o_1\ \ldots\ |S\ s_n\ |P\ r_n\ |O\ o_n$
- Position ID:
  - The same as the original position ID used in BART, is the index of the token in the flattened sequence
- Triple Role ID:
  - 3 values for a specific triple (s_i, r_i, o_i): 1 for the subject s_i, 2 for the relation r_i, and 3 for the object o_i.
- Tree level ID:
  - The distance (number of relations) from the root which is the source vertex of the RDF graph

Model	Origin	+ Position
BART-base	139.32M	139.43M
Distil-BART-xsum	305.51M	305.53M
BART-large	406.29M	406.31M
T5-base	222.88M	222.90M
T5-large	737.64M	737.65M

Experiment Dataset

Webnlg[1]

- The WebNLG Challenge dataset which contains 18,102/2,268/4,928 graph-description pairs for training, validation, and testing set respectfully.
- The input describes entities belonging to 15 distinct DBpedia categories Astronaut, University, Monument, Building, ComicsCharacter, Food, Airport, SportsTeam, WrittenWork, Athlete, Artist, City, MeanOfTransportation, CelestialBody, and Politician.

Wikipedia Pre-training Resource

- Based on Wikipedia dump and Wikidata crawled in March 2020 from 15 related categories in the WebNLG dataset
- For each Wikipedia article, query its corresponding WikiData triples and remove sentences which contain no values in the Wikidata triples to form graph-text pairs.
- Remove triples and description pairs that have already appeared in the WebNLG dataset.
- Obtain 542,192 data pairs.

salesforce

[1] Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. (2017, September). The webnlg challenge: Generating text from rdf data. In Proceedings of the 10th International Conference on Natural Language Generation (pp. 124-133).

Results for Various Pre-trained Models over All Categories on Development Set

Model	BLEU↑	P↑	R↑	F1↑
BART-base	57.8	68.7	68.9	67.0
+ Wikipedia	**59.7**	**69.6**	**70.7**	**68.4**
+ Position	58.8	68.7	69.9	67.6
+ Wiki + Position	57.3	67.8	69.0	66.6
BART-large	58.3	67.9	69.4	66.8
+ Wikipedia	59.0	68.0	**70.4**	**67.4**
+ Position	58.1	67.6	69.4	66.6
+ Wiki + Position	**60.0**	**68.6**	69.2	67.1
distill-BART-xsum	59.1	69.9	70.6	68.5
+ Wikipedia	59.8	69.7	**71.1**	**68.8**
+ Position	59.2	69.8	70.2	68.3
+ Wiki + Position	**59.9**	**70.1**	70.1	68.7
T5-base	**61.2**	72.3	72.0	70.6
+ Wikipedia	60.9	72.0	71.8	70.2
+ Position	60.8	**72.4**	**72.4**	**70.8**
+ Wiki + Position	60.3	72.2	72.0	70.5
T5-large	60.0	71.6	72.1	70.2
+ Wikipedia	61.3	72.2	72.0	70.5
+ Position	60.6	72.1	72.4	70.6
+ Wiki + Position	**61.9**	**72.8**	**73.5**	**71.6**

System Results on WebNLG Test Set with Official Scripts

Model		BLEU (%)↑			METEOR ↑			TER ↓		
		Seen	Unseen	All	Seen	Unseen	All	Seen	Unseen	All
Without Pretrained LM	Gardent et al. (2017)	54.52	33.27	45.13	0.41	0.33	0.37	0.40	0.55	0.47
	Moryossef et al. (2019)	53.30	33.31	37.24	0.44	0.34	0.39	0.47	0.56	0.51
	Zhao et al. (2020)	64.42	38.23	52.78	0.45	0.37	0.41	0.33	0.53	0.42
With Pretrained LM	Radev et al. (2020)	52.86	37.85	45.89	0.42	0.37	0.40	0.44	0.49	0.51
	Kale (2020)	63.90	52.80	57.10	**0.46**	0.41	**0.44**	-	-	-
	Riberiro et al. (2020)	64.71	53.67	59.70	**0.46**	**0.42**	**0.44**	-	-	-
Our Model	T5-large + position + Wiki	**66.07**	**54.05**	**60.56**	**0.46**	**0.42**	**0.44**	**0.32**	**0.41**	**0.36**

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In Proceedings of the 10th International Conference on Natural Language Generation, pp 124–133.
Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 pp 2267–2277.
Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi.
2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp 2481–2491
Linyong Nan, Dragomir Radev, Rui Zhang, Amrit
Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: Open-domain structured data record to text generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp 432–447
Mihir Kale. 2020. Text-to-text pre-training for data-totext tasks. Computation and Language Repository, arXiv:2005.10433.
Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schutze, ¨and Iryna Gurevych. 2020b. Investigating pretrained language models for graph-to-text generation. arXiv preprint arXiv:2007.08426

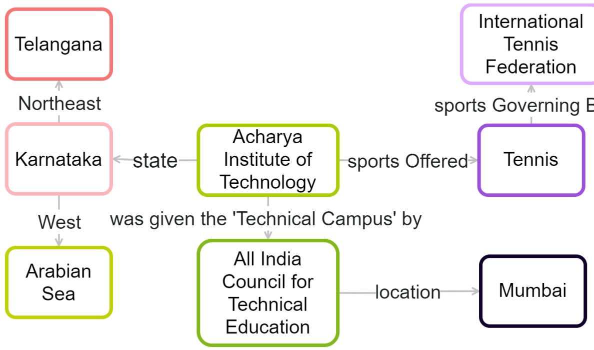
System Results on WebNLG Test Set with BERTScore

Model	**P↑**	**R↑**	**F1↑**
Gardent et al. (2017)	88.35	90.22	89.23
Moryossef et al. (2019)	85.77	89.34	87.46
Radev et al. (2020)	89.49	92.33	90.83
Riberiro et al. (2020)	98.36	91.96	90.59
T5-large + position + Wiki	**96.36**	**96.13**	**96.21**

Results of T5 Large

The state of Karnataka is located southwest of Telangana and east of the Arabian Sea. It is the location of the Acharya Institute of Technology which was granted the Technical Campus status by the All India Council for Technical Education in Mumbai. The Institute is affiliated with the *Visvesvaraya Technological University* and offers the sport of tennis. [International Tennis Federation]

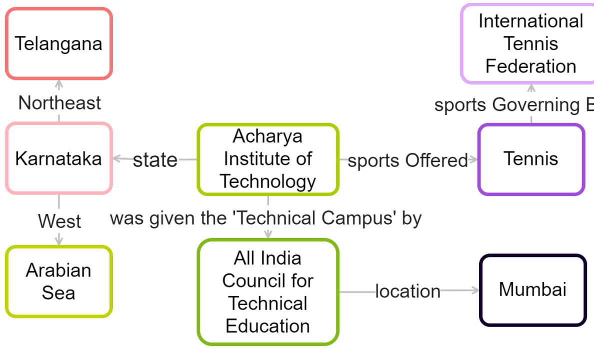
Problem: Hallucinate factual knowledge, missing facts



Results of T5 Large + Wiki

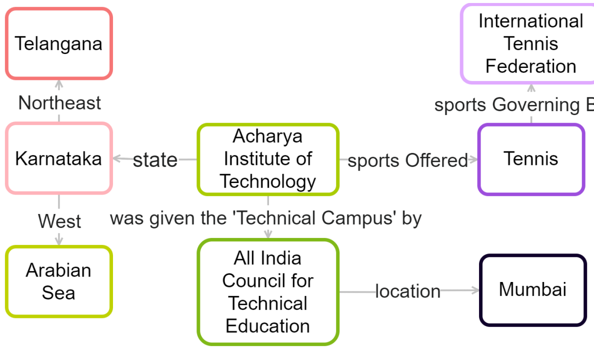
The Acharya Institute of Technology is located in the state of Karnataka. It was given the Technical Campus status by the All India Council for Technical Education which is located in Mumbai. The institute offers tennis **and has Telangana to its northeast and the Arabian Sea to its west.** [International Tennis Federation]

Problem: Missing facts, incorrect relations



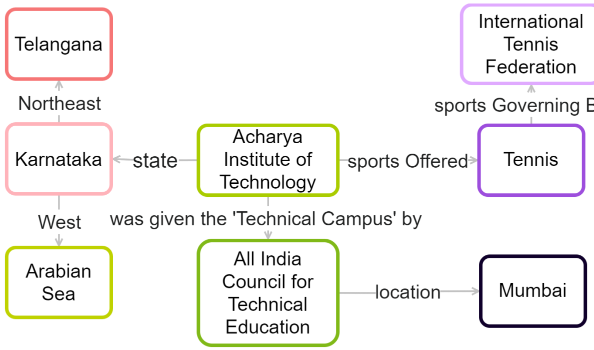
Results of T5 Large + Position

The Acharya Institute of Technology is located in the state of Karnataka which has Telangana to its northeast and the Arabian Sea to its west. It was given the Technical Campus status by the All India Council for Technical Education in Mumbai. The Institute offers tennis which is governed by the International Tennis Federation.



Results of T5 Large + Wiki + Position

The Acharya Institute of Technology is located in the state of Karnataka which has Telangana to its northeast and the Arabian Sea to its west. The Institute was given the Technical Campus status by the All India Council for Technical Education in Mumbai. One of the sports offered at the Institute is tennis which is governed by the International Tennis Federation.



Impact of Wikipedia Fine-tuning

- Capture unseen relations:
  - *(Olusegun Obasanjo, in Office While Vice President, Atiku Abubakar)* is translated to *"His vice president is Atiku Abubakar"*
  - *(Aaron Turner, active Years Start Year , 1995)* is translated to *"started playing in 1995"*.
- Combines relations with the same type together with correct order:
  - Given two death places of a person such as *(Alfred Garth Jones, deathplace, Sidcup)* and *(Alfred Garth Jones, deathplace, London)*, the model generates: *"died in Sidcup, London"* instead of generating two sentences or placing the city name ahead of the area name.

Impact of Position Embeddings

- Reduce the errors introduced by pronoun ambiguity:
  - For a KG which has *“leader Name”* relation to both country’s leader and university’s dean, position embeddings can distinguish these two relations by stating *“Denmark’s leader is Lars Løkke Rasmussen”* instead of *“its leader is Lars Løkke Rasmussen”*.
- Arrange multiple triples into one sentence:
  - Combining the city, the country, the affiliation, and the affiliation’s headquarter of a university into a single sentence: *“The School of Business and Social Sciences at the Aarhus University in Aarhus, Denmark is affiliated to the European University Association in Brussels”*.

Remaining Challenges

- Biased against the occurrence of patterns that would enable it to infer the right relation:
  - Confuse *"active Years Start Year"* relation with the birth year.
- Fail to capture the deep connections between the subject and the object:
  - For the relation *"doctoral Student"*, the model mistakenly considers a professor as a Ph.D. student
  - Treat an asteroid as a person because of its epoch date.
- Miss relations for complex graph input
  - For a soccer player with multiple clubs, the system might confuse the subject of one club's relation with another club.

The programs, data and resources are publicly available for research purpose at:
https://github.com/ EagleW/Stage-wise-Fine-tuning

Codes and Data
are public at: