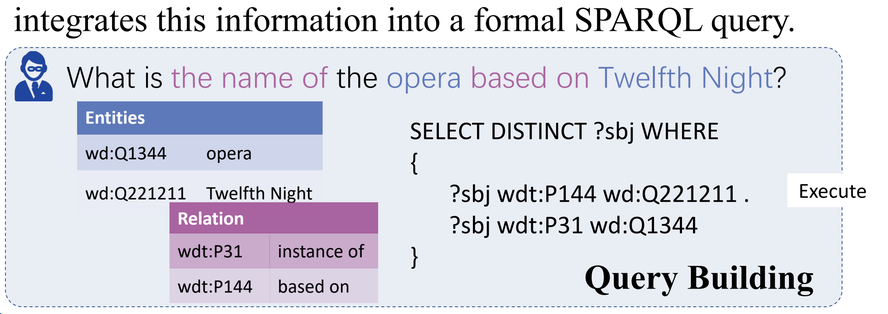
Enhancing SPARQL Generation by Triplet-order-sensitive Pre-training

Chang Su[1], Jiexing Qi[1], He Yan[2], Kai Zou[3], Zhouhan Lin[1]*

[1]Shanghai Jiao Tong University;
[2]ProtagoLabs Inc.;
[3]NetMind.AI LTD;
* corresponding author; our implementation is available at https://github.com/LUMIA-Group/TosT5.

Task: Query Building in KGQA

Given a question posed in natural language, the Knowledge Graph Question Answering (KGQA) system's objective is to retrieve the correct answer from the KG. A standard KGQA system typically involves three main steps: 1) Entity Linking (EL), 2) Relation Linking (RL), and 3) **Query Building (QB)**. Once the entities and relations are linked, the query-building module integrates this information into a formal SPARQL query.



Limitation of existing works

Existing works mainly focus on improving syntax correctness in the SPARQL generation task. However, errors like **triplet-flip errors**, which do not compromise the grammatical correctness or executability of queries but yield incorrect or unavailable results, cannot be resolved through the schema-level constraints proposed by these works.

Target: select distinct ?ans where { ?ans wdt:P279 wd:Q25356}

Wrongly predicted: select distinct ?ans where { wd:Q25356 wdt:P279 ?ans}

where wdt:P279 and wd:Q25356 are the IRIs for relation subclass of and entity Populus, and the positions of e_s and e_o are mistakenly flipped. Real-world scenarios are more complex.

Method

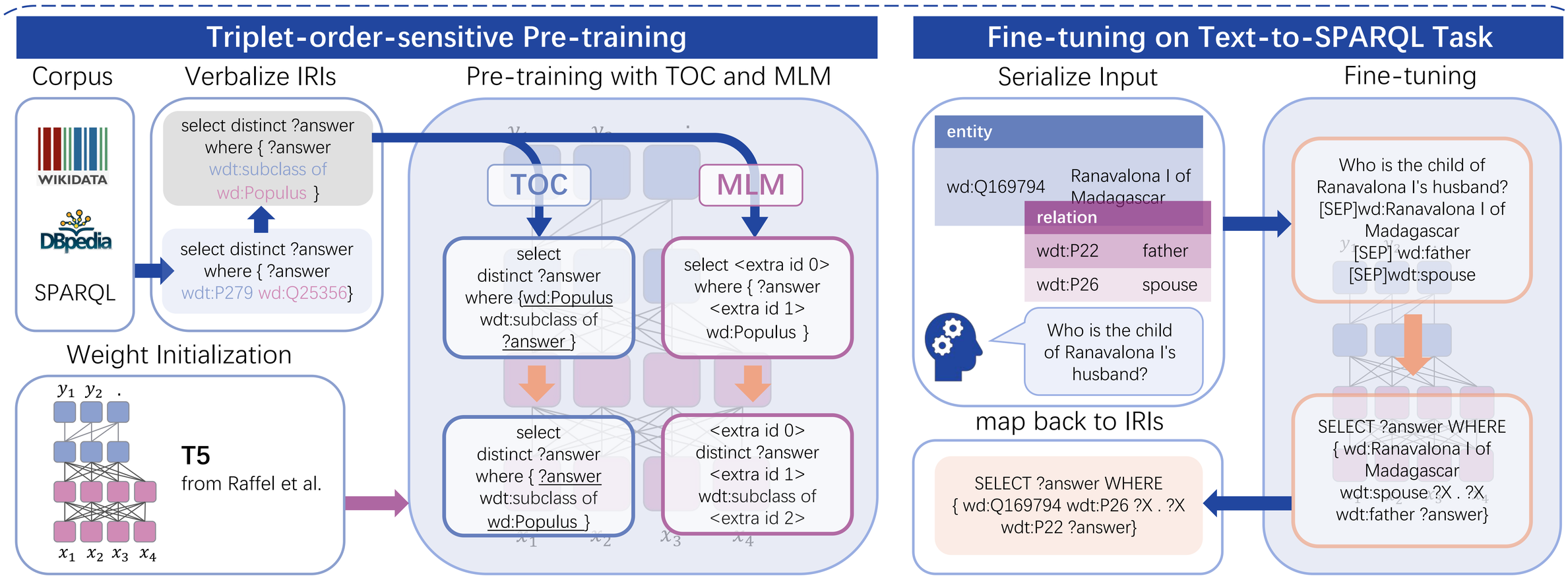


Figure 1. The overview of our approach. The TosT5 model first undergoes the triplet-order-sensitive pre-training stage and then is fine-tuned on the downstream task.

We introduce the TosT5 model, using T5 as the backbone and **a novel pre-training stage** to address triplet-flip errors. This pre-training employs **two objectives** to enhance the model's understanding of SPARQL. **Verbalizing IRIs** is proposed to help the model better understand. Following the pre-training, our TosT5 model is fine-tuned on the Text-to-SPARQL task.

- **Triplet Order Correction (TOC)**

The triplet (e_s, r, e_o) can be permuted into any arrangement. For multiple triplets, shuffling is independent. The model predicts the initial order and reconstructs the original SPARQL statements. Given input $\tilde{x}$, it aims to generate x, with the TOC loss defined as:

$$\mathcal{L}_{\text{TOC}} = -\sum_{i=1}^{|x|} \log P_\Theta(x_i | x_{<i}; \tilde{x})$$

- **Masked Language Modeling (MLM)**

We employ a span masking strategy with sentinel tokens, forming output y by concatenating these with real masked tokens. The training loss is denoted as:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i=1}^{|y|} \log P_\Theta(y_i | y_{<i}; x)$$

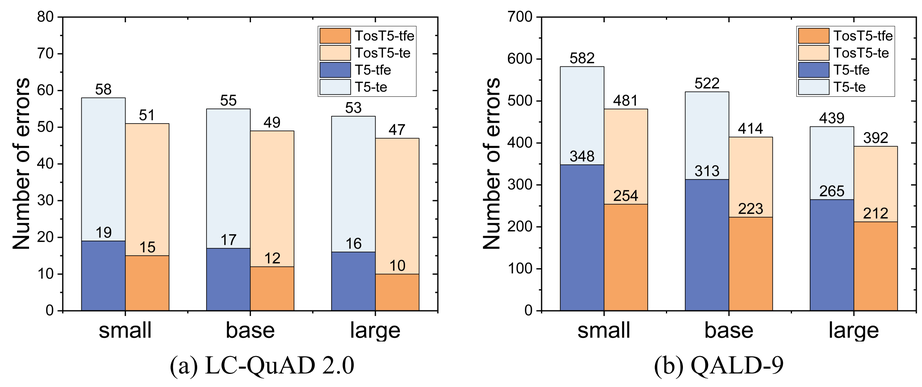
Result

Table 1. Experimental results for LC-QuAD 2.0. † denotes only gold entities are provided. ⋆ denotes our re-implemented results. The values in parentheses indicate the 95% confidence intervals. The best performance for each T5 size is bolded.

Approach	F1	QM
AQG-net	44.9	37.4
Multi-hop QGG	52.6	43.2
CLC+BERT	59.3	55.4
BART	64.0	-
PGN-BERT	77.0	-
PGN-BERT-BERT	86.0	-
$\text{SGPT}_{Q,K}$ †	89.0	-
T5-base	91.0	-
T5-small⋆	92.4	90.2
T5-base⋆	93.4	91.4
T5-large⋆	94.1	92.1
TosT5-small	**94.1**(±0.22)	**92.0**(±0.20)
TosT5-base	**95.0**(±0.27)	**93.2**(±0.24)
TosT5-large	**95.4**(±0.21)	**93.5**(±0.21)
T5-large†	85.3	74.9
TosT5-large†	90.3(±0.31)	79.2(±0.27)

Table 2. Experimental results on QALD-9 and QALD-10 test set. The improvements compared to baselines are bolded.

Approach	QALD-9		QALD-10	
	QM	F1	QM	F1
TeBaQA†	-	28.8	-	-
$\text{SGPT}_{Q,K}$ †	-	67.8	-	-
T5-small	55.1	64.4	33.9	38.8
T5-base	58.1	69.5	36.4	39.8
T5-large	59.6	71.6	35.7	45.3
TosT5-small	61.1(**+6.0**)	72.8(**+8.4**)	40.1(**+6.2**)	47.2(**+8.4**)
TosT5-base	61.8(**+3.7**)	75.8(**+6.3**)	39.1(**+2.7**)	51.4(**+11.6**)
TosT5-large	63.2(**+3.6**)	73.2(**+1.6**)	39.0(**+3.3**)	51.2(**+5.9**)



(a) LC-QuAD 2.0 (b) QALD-9

Figure 2. Error analysis on models' prediction. "tfe" is short for "triplet-flip error", and "te" is short for "triplet error".

- Our method achieves **state-of-the-art performances** on three widely-used benchmarks: LC-QuAD 2.0, QALD-9 and QALD-10.
- We perform a thorough error analysis showing that our method effectively **reduces triplet-flip errors**, indicating that pre-training enhances the model's understanding of triplet element order.

Conclusion:

In this paper, we introduce a triplet-order-sensitive T5 model, namely TosT5, specifically designed to address the triplet-flip errors exhibited in the Text-to-SPARQL task. We adopt T5 as the backbone and utilize two pre-training objectives to enhance the model's comprehension of the SPARQL language. We also propose to verbalize IRIs during training to better leverage the pre-trained model's ability in language understanding. After undergoing the pre-training, our TosT5 model is fine-tuned on the downstream Text-to-SPARQL task, achieving new state-of-the-art performances on three well-known KGQA datasets. We believe our model will help enhance the overall performance of the KGQA system.