

DipInfo-UniTo at the GEM'24 Data-to-Text Task: Augmenting LLMs with the Split-Generate-Aggregate Pipeline

Michael Oliverio ✧, Pier Felice Balestrucci ✧, Alessandro Mazzei ✧, Valerio Basile ✧

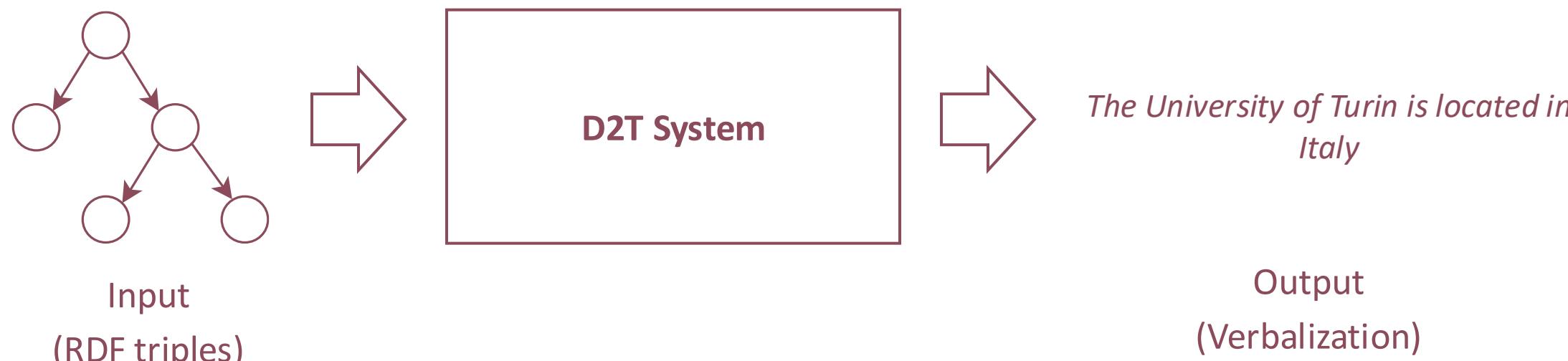
✧ Univeristy of Turin, Computer Science
(Italy)



GitHub repository

GEM'24 D2T Shared Task

- Challenge on RDF-to-Text generation between different teams
- Based on 2 subtask:
 - WebNLG-based (D2T-1)
 - Wikidata-based (D2T-2)
- For each subtask, there are 3 parallel datasets:
 - Factual (FA): extracted directly from the source
 - Counterfactual (CFA): with swapped subject-object
 - Fictional (FI): with entities generated by LLM



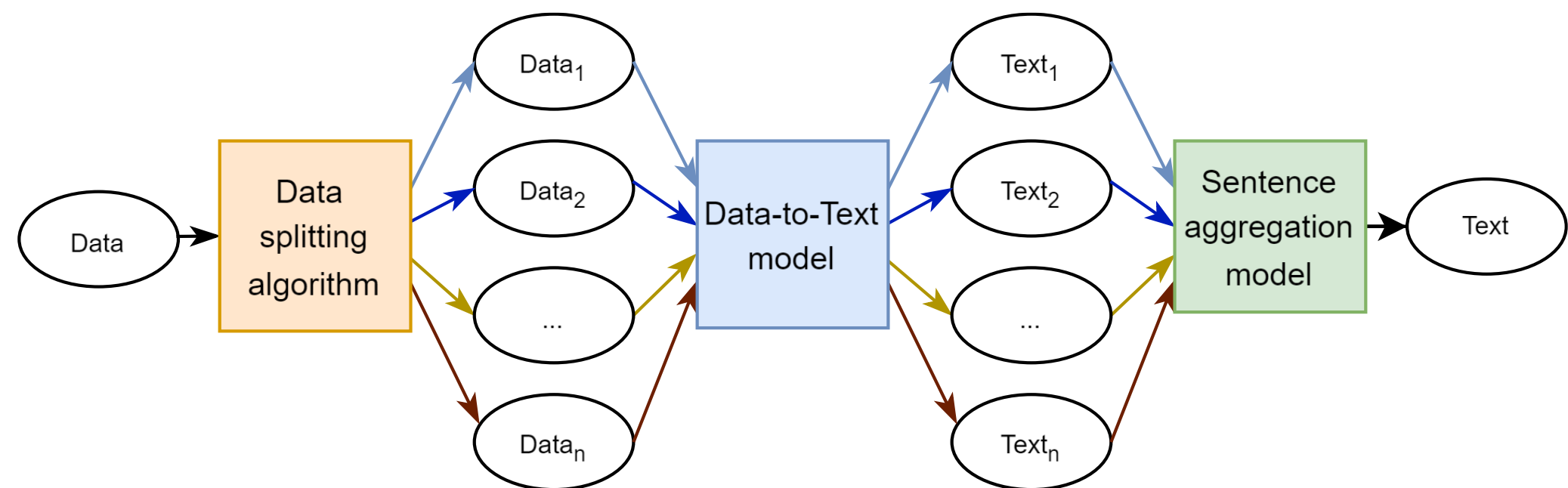
The main idea

Questions

- What is the main problem with LLM for D2T task?
→ Presence of hallucinations in the generations
- What does their presence mean?
→ Lack of adequacy
- When this problem occur most?
→ In presence of complicated input
- How to avoid or reduce this problem?
→ Splitting the input into subsets of simpler input

Our proposal

- Create a pipeline called SGA (split-generate-aggregate), to simplify the generation into sub-generations and aggregate them into a single text



Metodology

- We create a system based on 3 steps:
 - Data splitting
 - Generation
 - Aggregation

Data splitting

- It is based on an algorithm that splits the data unit into subsets, each containing at least at least three triples
- To perform the split, the relationships between the triples, which can be either chain or sibling (CITAZIONE), are identified within the data units



Generation

- Step performed using Mistral-7B (CITAZIONE)
- Fine-tuned using WebNLG 3.0 (CITAZIONE)
- Only 20% of the dataset are used for fine-tuning due our low computational resources
- QLoRA quantization technique (Dettmers et al., 2023 CITAZIONE) used to simplify the fine-tuning process and reduce the computational impact

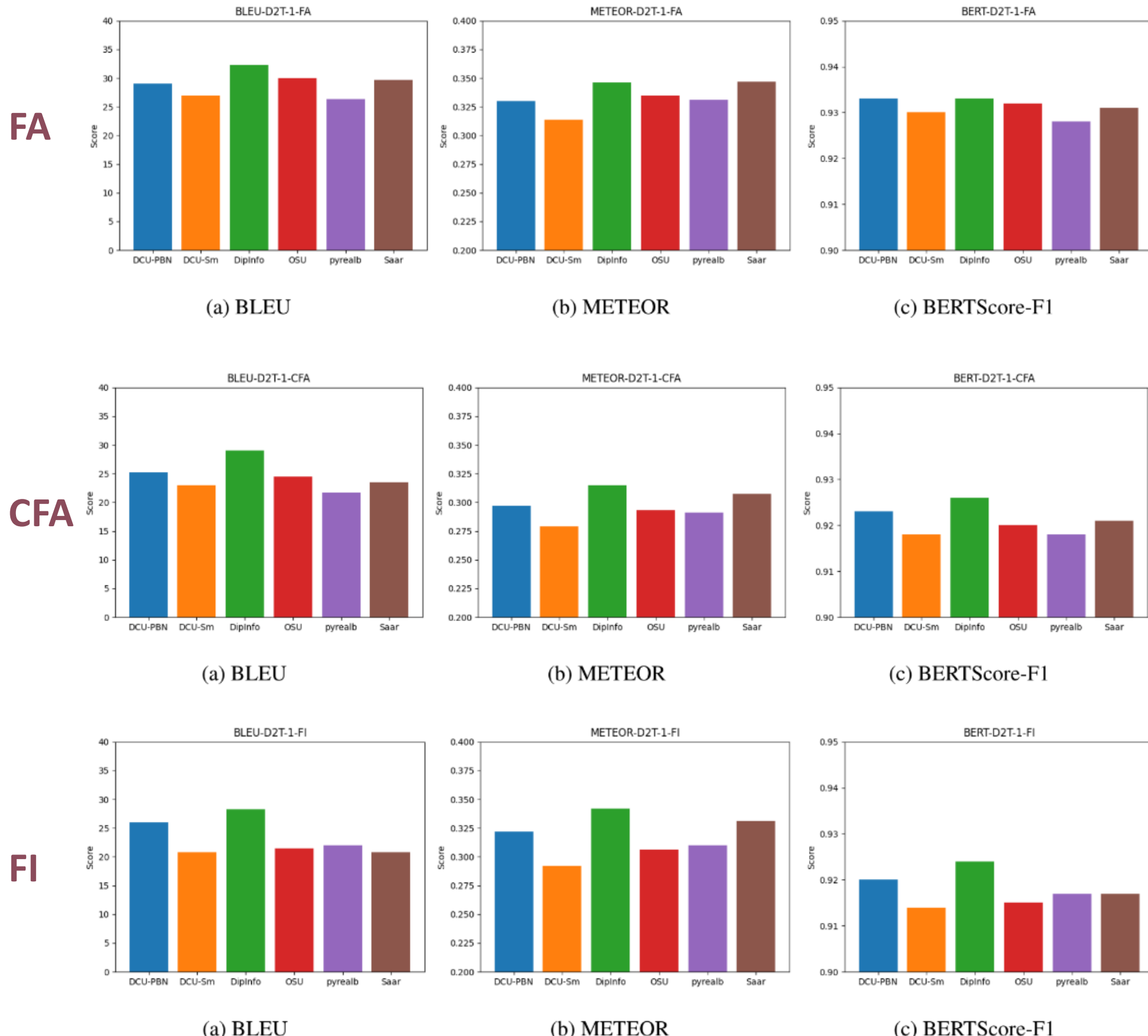
Aggregation

- Step based on Mistral-7B
- In this case we use a zero-shot prompting technique for the task of aggregation

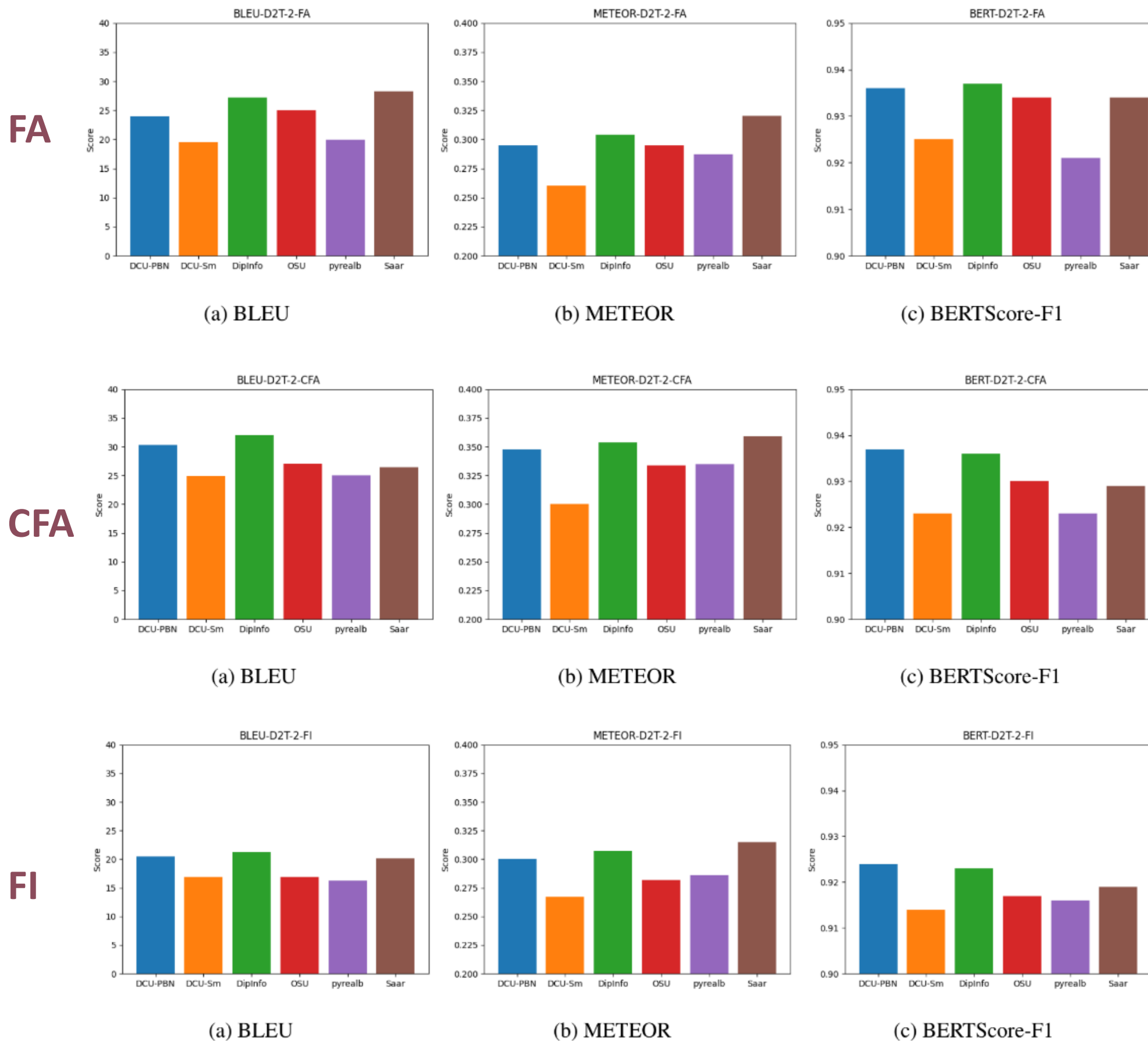
Instruction=“You have to aggregate and paraphrase together the following sentences. You have to generate the result in English”

Results

Subtask 1



Subtask 2



Considerations

- DipInfo-UniTo** ranked first in half of the trials and secured second place in the remaining experimentsIt
- This results demonstrates that it is a robust system capable of producing high-quality inferences

Future improvements

- Develop a more sophisticated data splitting algorithm
- Utilize the entire training set of WebNLG 3.0 to fine-tune the LLM for the D2T task
- Dive deeper into the aggregation task, aiming to enhance aggregation results