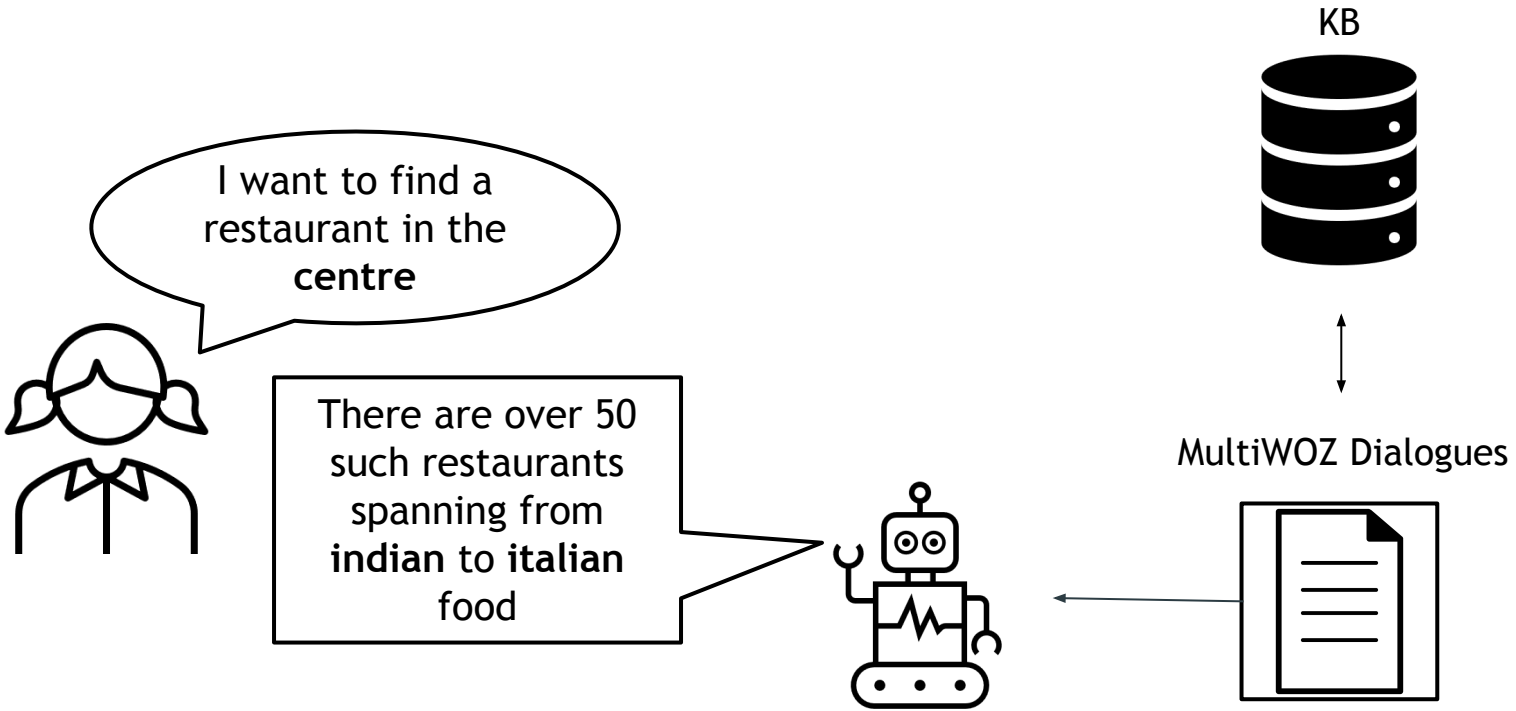
Towards Cost-effective Multi-style Conversations: A Pilot Study in Task-oriented Dialogue Generation

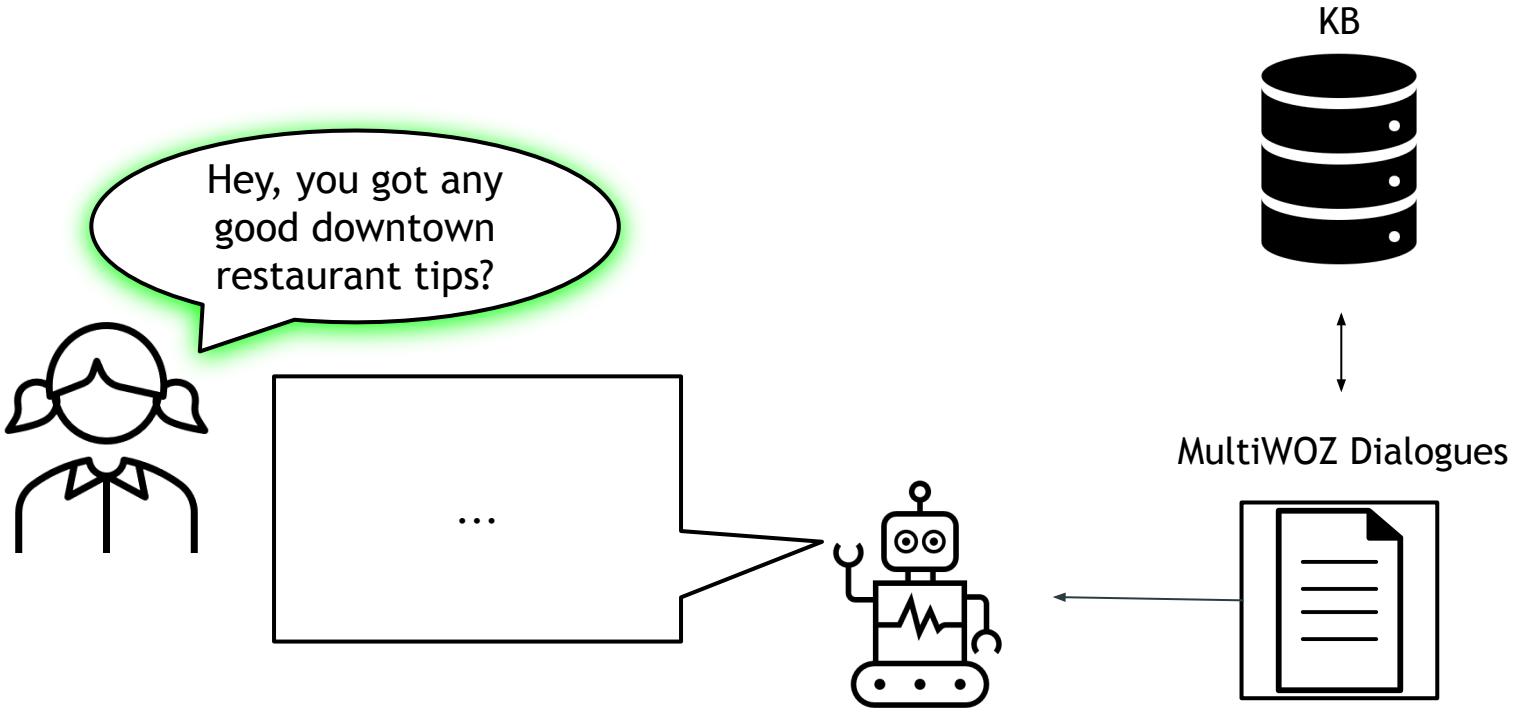
Tiziano Labruna
Fondazione Bruno Kessler, Trento
Free University of Bolzano-Bozen
tlabruna@fbk.eu

Bernardo Magnini
Fondazione Bruno Kessler, Trento
magnini@fbk.eu

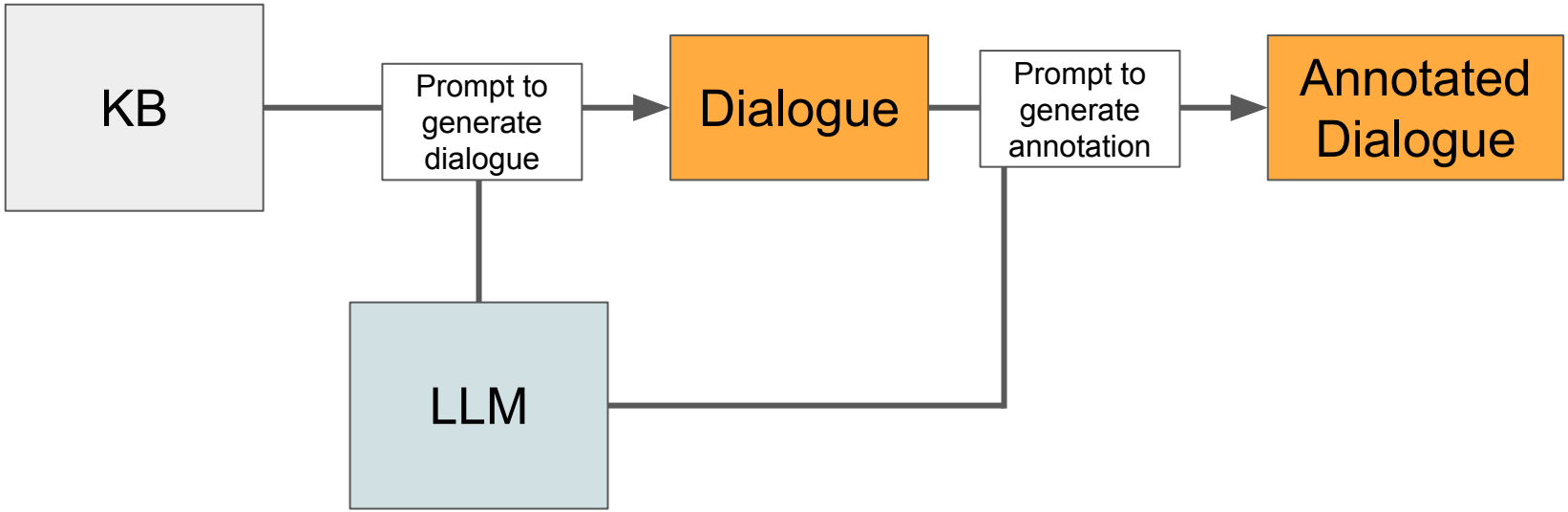
Task-oriented Dialogue Systems



Task-oriented Dialogue Systems - Style changes

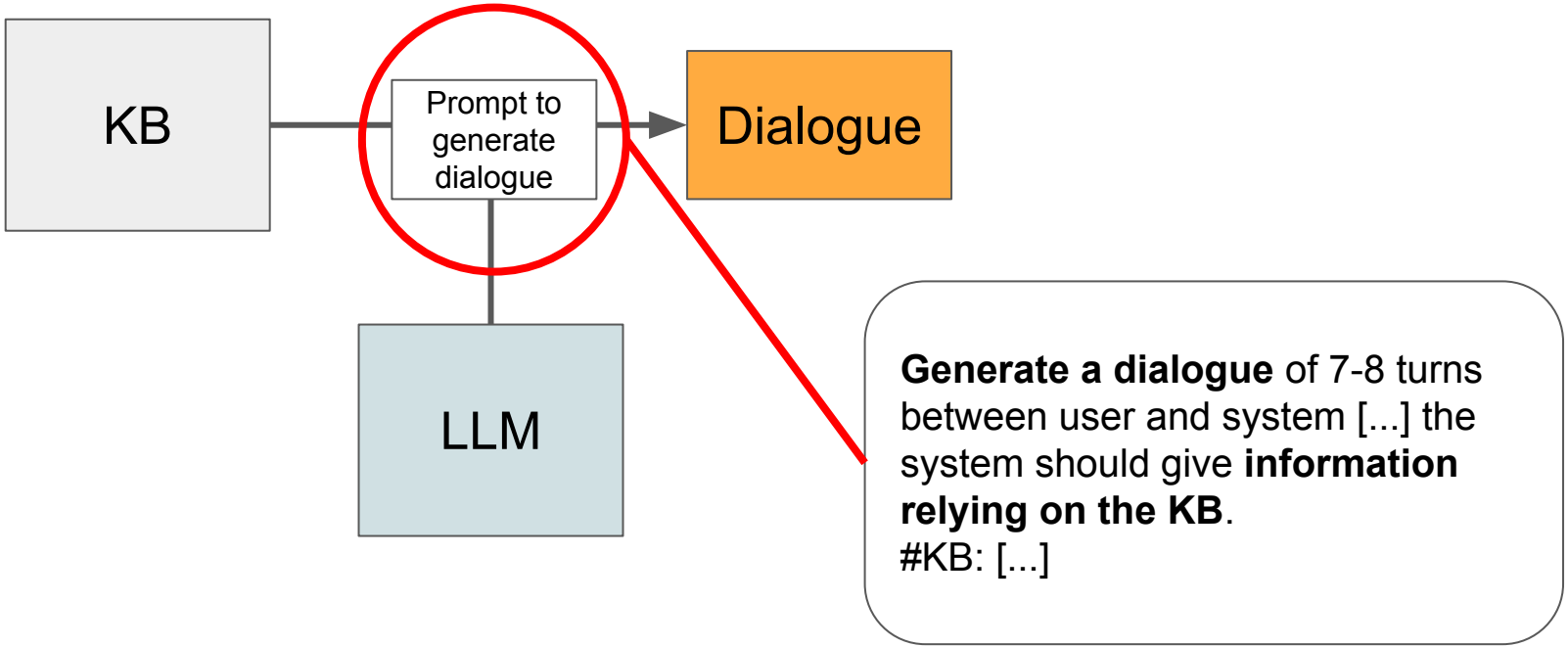


Task-oriented Dialogue Generation

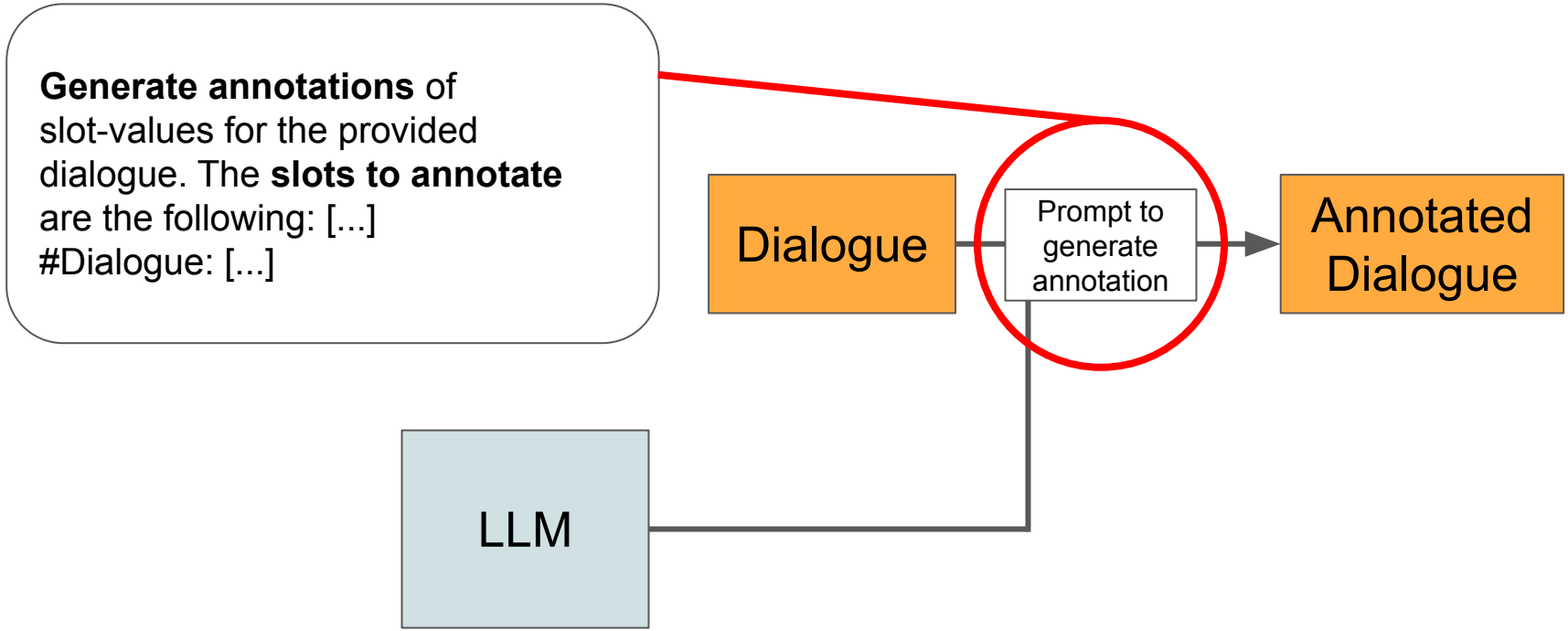


Domain Knowledge Base

Task-oriented Dialogue Generation



Task-oriented Dialogue Generation



Multi-style Dialogue Resources

MultiWOZ

- **Task-oriented** dialogues collected **manually** through WOZ
- **1,180** dialogues as training set
- **131** dialogues as test set

NeutralGPT

- Generated with **gpt-3.5-turbo**, prompted to produce a **neutral** style
- **1,180** dialogues as training set
- **131** dialogues as test set

MultistyleGPT

- **Merge** of half of **MultiWOZ** and half of **NeutralGPT**
- **1,180** dialogues as training set
- **131** dialogues as test set

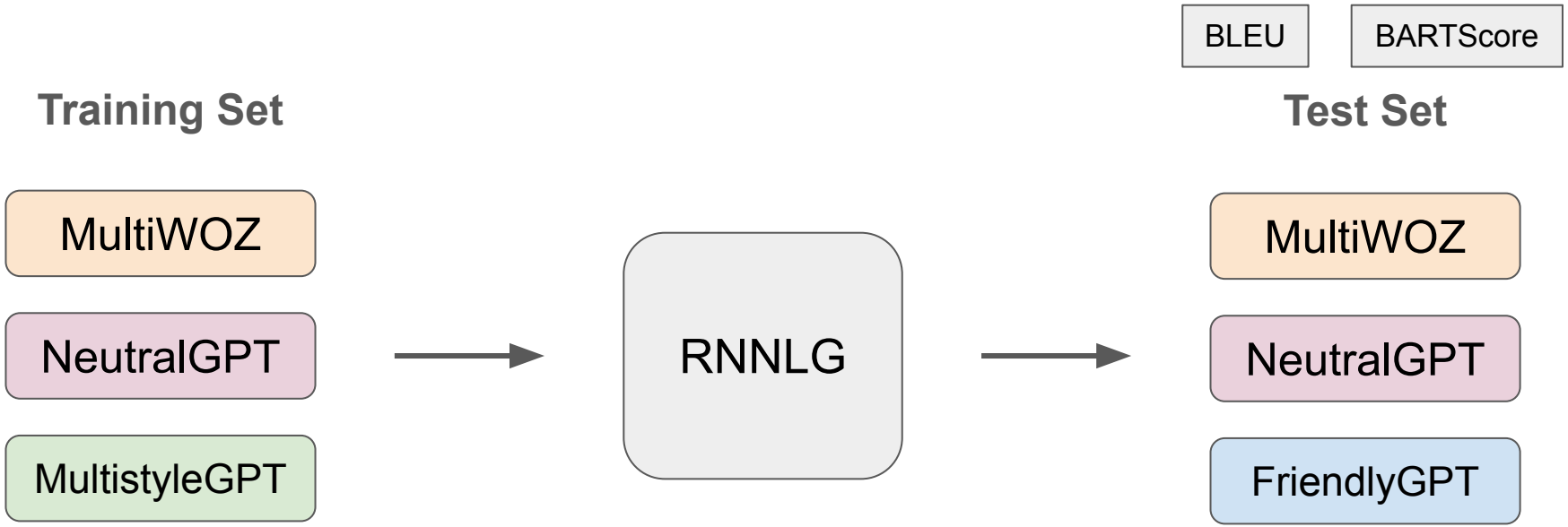
FriendlyGPT

- Generated with **gpt-3.5-turbo**, prompted to produce a **friendly** style
- **131** dialogues as test set

Multi-style Dialogue Resources

Dataset Characteristic	MULTIWOZ	NEUTRALGPT	FRIENDLYGPT	MULTISTYLEGPT
Avg. System turns per dialogue	4.39	6.05	7.28	5.32
Avg. Turn length	16.27	25.06	19.95	21.09
Avg. Slots per turn	2.56	3.02	1.96	2.73
Tot. Unique slot-values	443	439	400	474
Avg. Intents per turn	1.54	1.30	1.35	1.35
Avg. Utterances per turn	1.8	2.51	2.40	2.20
Type-Token Ratio	0.10	0.05	0.09	0.08

Experimental Setting



Experiment Results

Training-set	Test-set	BLEU	BARTScore
MULTIWOZ	MULTIWOZ	0.437	-4.388
NEUTRALGPT	MULTIWOZ	0.088	-5.465
MULTISTYLEGPT	MULTIWOZ	0.340	-4.443
MULTIWOZ	NEUTRALGPT	0.184	-4.841
NEUTRALGPT	NEUTRALGPT	0.365	-4.422
MULTISTYLEGPT	NEUTRALGPT	0.343	-4.736
MULTIWOZ	FRIENDLYGPT	0.122	-4.991
NEUTRALGPT	FRIENDLYGPT	0.181	-4.616
MULTISTYLEGPT	FRIENDLYGPT	0.199	-4.665

Experiment Results

- **Training** an NLG model on **one style** and **evaluating** it on a **different style** results in **low performance**.
- **Training** an NLG model on a **multi-style** dataset results in **better performance** than training on a single style.
- The **style** of the training set has a **strong impact** on the performance of a **dialogue system**.

Conclusions

- We presented a **cost-effective** methodology for **generating diverse conversational style** datasets.
- **Single style** models struggle when exposed to different styles.
- Training a model on **multiple styles** can improve performance.
- Future research should **further evaluate** how integrating **different styles** in the training can enhance performance.

Thank you

Tiziano Labruna[1,2] – tlabruna@fbk.eu

[1] Fondazione Bruno Kessler

[2] Free University of Bolzano-Bozen