Project Proposal

Multimodal Reference Resolution

March 28, 2022

Student: Alejandro Santorum Varela

Supervisors: Dr. Svetlana Stoyanchev and Dr. Kate Knill

UNIVERSITY OF CAMBRIDGE
Department of Engineering

UNIVERSITY OF
CAMBRIDGE

Multimodal Reference Resolution proposal

Project proposal

Reference Resolution in a dialogue system is the task of identifying the object referred by the dialogue speaker. A system is multimodal if both visual and dialogue data are available to identify the object. It is useful to successfully address this problem since it can have many applications, such as human-robot communication systems. Human-robot dialog systems allow a user to give instructions to a robot to complete tasks (navigation, moving objects, etc.). In this case, the robot would process user's dialog and the current scene to determine which object the user is referring to.

This project aims to tackle this problem, following the second track of the Tenth Dialog System Technology Challenge (DSTC10) [3], that is the tenth edition of an initiative to provide a common testbed for the task of Dialog State Tracking. The proposed tracks for the 2021 edition were: multimodal disambiguation, multimodal coreference resolution, multimodal dialog state tracking and multimodal dialog response generation and retrieval. The provided dataset for this track is the SIMMC2 dataset [2] created by Meta's Research team [1].

We are going to study and review the published solutions for the challenge, like the UNITER-based solution [4] of team 7, that ranked 2nd in the DSTC10 with a test-std F1 score of 73.3%. The top performing team (team 4) formed by KAIST, ETRI and Samsung Research [5] achieved a final test-std F1 score of 75.8% using a BART transformer model [6].

The ultimate goal consists in improving the top performance of 75.8% investigating and modifying the current solutions, as well as building new transformer models (GPT2, GPT3, BART, BERT, etc.). Developing a research publication is a possible option if the project is a complete success.

Workplan

The project development will not fully start until the end of April. Until then, the dataset [1] and the DSTC10 [3] will be studied in parallel with the exam period.

After the exam period, weekly meetings will be set and we will review the literature, focusing in the current solutions ([4], [5]). In the second half part of May and in the first half part of June the UNITER-based solution [4] and the BART-based top performing solution [5] will be implemented and their limitations examined.

A week in mid June is going to be used to elaborate a poster for the Research Review Day. Gathering the results and representing them graphically are going to be the priorities. After that, previous solutions will be further investigated.

July will consists almost entirely in improving the current models, looking for beating the best performing systems.

In August, the possible final refinements are executed and the report write-up will be carried out.

This workplan is further described below:

- From **March 28th** to **April 25th**: Data exploration and challenge review.
- From **April 26th** to **May 15th**: Literature review.
- From **May 16th** to **June 12th**: Study and implement UNITER-based solution [4].
- From **June 13th** to **June 20th**: Poster preparation for Research Review day.
- From **June 21st** to **July 3rd**: Study and implement other solutions, such as team 4 proposal [5].
- From **July 4th** to **July 31st**: Modify existing models or investigate new ones to improve top performance.
- From **August 1st** to **August 18th**: Report write-up.

Resource declaration

- **Resources**: Toshiba Europe Ltd. is providing advanced computing resources, such as a laptop and access to a CPU/GPU cluster. MLSALT computing resources might also be used.
- **Data**: The project will use the SIMMC2 dataset [2], published by Meta's Research team [1].
- **Human participants**: The project does *not* involve studies with human participants.

Bibliography

[1] SIMMC2. Meta's Research team. GitHub. https://github.com/facebookresearch/simmc2/tree/main/.

[2] ”SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations”. Kotur et.al. *Computing Research Repository (CoRR)*. 2021. https://aclanthology.org/2021.emnlp-main.401.pdf.

[3] The Tenth Dialog System Technology Challenge (DSTC10). 2021. https://sites.google.com/dstc.community/dstc10/home

[4] ”UNITER-Based Situated Coreference Resolution with Rich Multimodal Input”. Yichen Huang, Yuchen Wang, Yik-Cheung Tam. *Computing Research Repository (CoRR)*. 2021. https://arxiv.org/abs/2112.03521.

[5] KAIST, ETRI and Samsung Research submission for DSTC10. GitHub. 2021. https://github.com/KAIST-AILab/DSTC10-SIMMC.

[6] ”Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. *Computing Research Repository (CoRR)*. Mike Lewis, Yinhan Liu, et.al. 2019. https://arxiv.org/pdf/1910.13461.pdf