Project Proposal

Multimodal Reference Resolution

April 2, 2022

Student: Alejandro Santorum Varela

Supervisors: Dr. Svetlana Stoyanchev and Dr. Kate Knill

UNIVERSITY OF CAMBRIDGE
Department of Engineering

Alejandro Santorum Varela

Trumpington Street
Cambridge CB2 1PZ
Tlf.: +44 01223 332600
http://www.eng.cam.ac.uk/

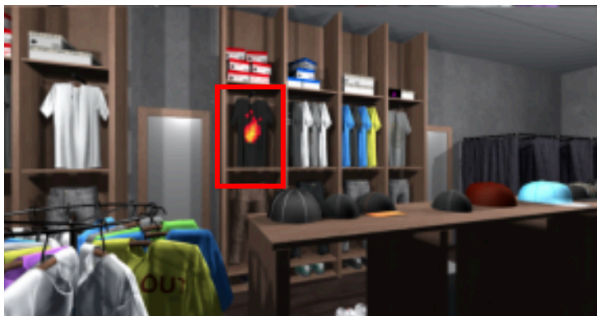
UNIVERSITY OF CAMBRIDGE

Multimodal Reference Resolution proposal

Project proposal

Dialogue systems are natural language interfaces that allow users to perform tasks, such as database search or command and control, using speech. Multimodal dialogue systems use vision and speech for communication with the user. Examples of multimodal dialogue systems include human-robot interaction, virtual agents, etc. Reference resolution in a multimodal dialogue system involves identification of an object in a visual scene that the user refers to. Multimodal reference resolution is essential for interpreting user utterance in a multimodal dialogue system. For example, Human-robot communication 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.

Given a scene represented by an image and user utterances that contains a reference to an object in the scene, the task is to identify the referent objects. As an example, Figure 1 illustrates two scenes with several utterances that would describe the selected object.



(a) The user may refer to the selected object as: "the black t-shirt with the fire design", "the black t-shirt in the middle", "the t-shirt hanging below the red boxes", "the t-shirt to the right of the mirror", etc.

(b) The user may refer to the selected object as: "the yellow t-shirt", "the t-shirt to the left of the grey t-shirts", "the t-shirt hanging below the white boxes", "the yellow t-shirt next to the blue ones", etc.

Figure 1: Examples of the visual scenes.

This project aims to tackle multimodal reference resolution problem, which was recently addressed in one of the tracks of the Tenth Dialog System Technology Challenge (DSTC10) [3]. The 2021 competition included five tracks. Multimodal Coreference Resolution was one sub-task of the third track "SIMMC 2.0: Situated Interactive Multimodal Conversational AI" [2]. The provided dataset for this task is the SIMMC2 dataset [2] created by Meta's Research team [1]. The dataset consists of the simulated dialogues where the user interacts with an assistant to obtain recommendations for a piece of fur-

niture or a clothing item. The dialogues were further manually edited by crowdsoursing, resulting in 11244 dialogues, consisting of 117236 utterances and 1566 scene images in total. The dataset also includes metadata for each object referred in the dialog data. Each item in a catalog metadata has a unique `object_id`. Each `scene_json` defines the mapping from the `local_idx` (local to each dialog), to its canonical `object_id` reference, for each dialog.

In the DSTC10, the performance of the proposed models is assessed by F1 score, that is the harmonic mean of precision and recall.

We are going to study and review the published solutions for the challenge, including 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 goal of the project is to assess strengths and weaknesses of the previous approaches to multimodal reference resolution task and to design an approach that further improves the performance. We will explore the use of transformer model architectures (GPT2, GPT3, BART, BERT, etc.) and the possibility of incorporating the domain knowledge into the neural framework. Depending on the outcomes of the project, the results may also be written up in a research paper for a conference.

Workplan

The project development will start on the 25th 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 week of June both the UNITER-based solution [4] and the BART-based top performing solution [5] will be studied, and one of them will be selected to be replicated. After that, the weaknesses of the implemented existing model will be analyzed and improvements will be proposed.

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, the proposed improved model will be implemented.

In July we will be evaluating and assessing the proposed model, and improving it even further.

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

This workplan is further described below:

- March 28 - April 25: Data exploration and challenge review.
- April 25 - May 6 (2 weeks): Literature review on dialogue, coreference resolution.
- May 9 - May 20 (2 weeks): Review and study the existing methods ([4], [5]).
- May 23 - June 4 (2 weeks): Implement one of the existing methods.
- June 6 - June 11 (1 week): Analyze weaknesses and propose improvements.
- June 13th - June 20th (1 week): Poster preparation for Research Review day.
- June 21 - July 2 (2 weeks): Implement the proposed approach.
- July 4 - July 16 (2 weeks): Evaluation and assessment.
- July 8 - July 30 (2 weeks): Further improvement to the method.
- August 1 - 18 (2.5 weeks): 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