Using a cooperative Image-Wordle game to investigate collaborative dialogical decision making

TAMARA ATANASOSKA

SUPERVISED BY:
DR. JANA GÖTZE
PROF. DR. DAVID SCHLANGEN

Abstract

The field of explainability, or explainable AI(XAI), has been the focus of many research efforts in recent years. One way of providing insight into the decision-making of a model in a way that will be immediately interpretable to humans would be the option to place queries and receive answers in natural language. Agreement games offer an effective method to externalise the reasoning and decision-making process in human dialogues. This research fits in the broader assumption that it will be possible to train a model on data collected by playing these games to generate explanations for its decisions. **The project aims to increase the diversity and quality of the data collected by playing a cooperative image-grounded Wordle game. The main focus is creating an automated supply of image-term pairs characterisable by difficulty. Further, the project determines if those pairs elicit the correct type of dialogue and if those, or parts of them, can be used to train a model in the future.** The preliminary data is promising, and the data collection through the Amazon Mechanical Turk is still pending.

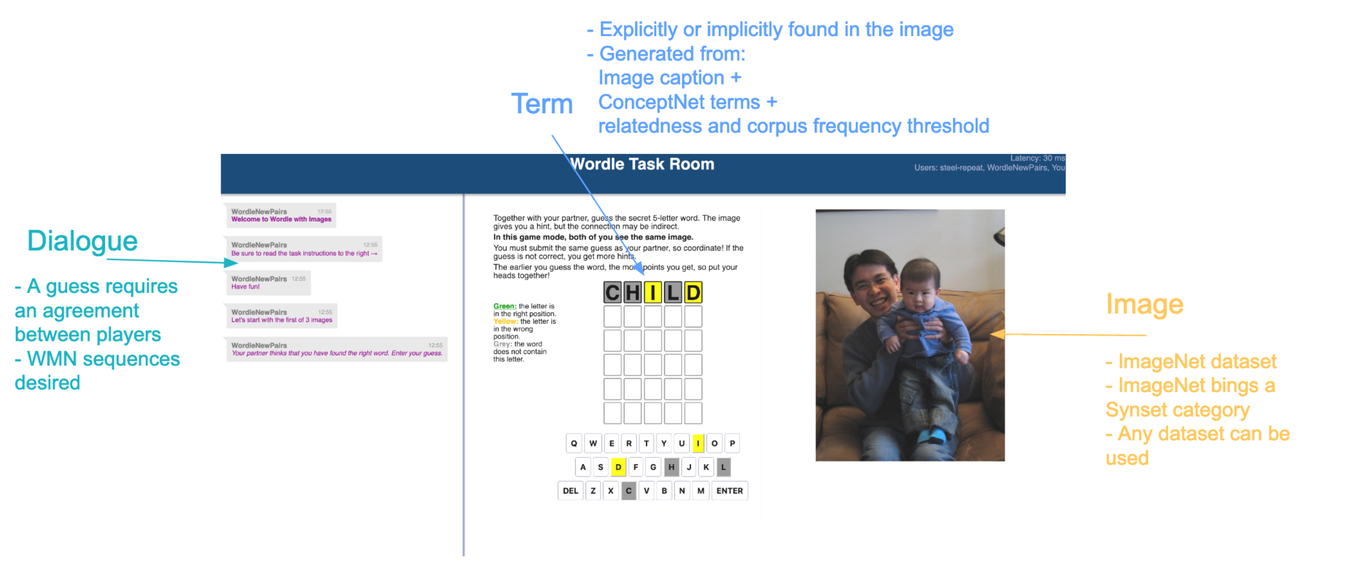
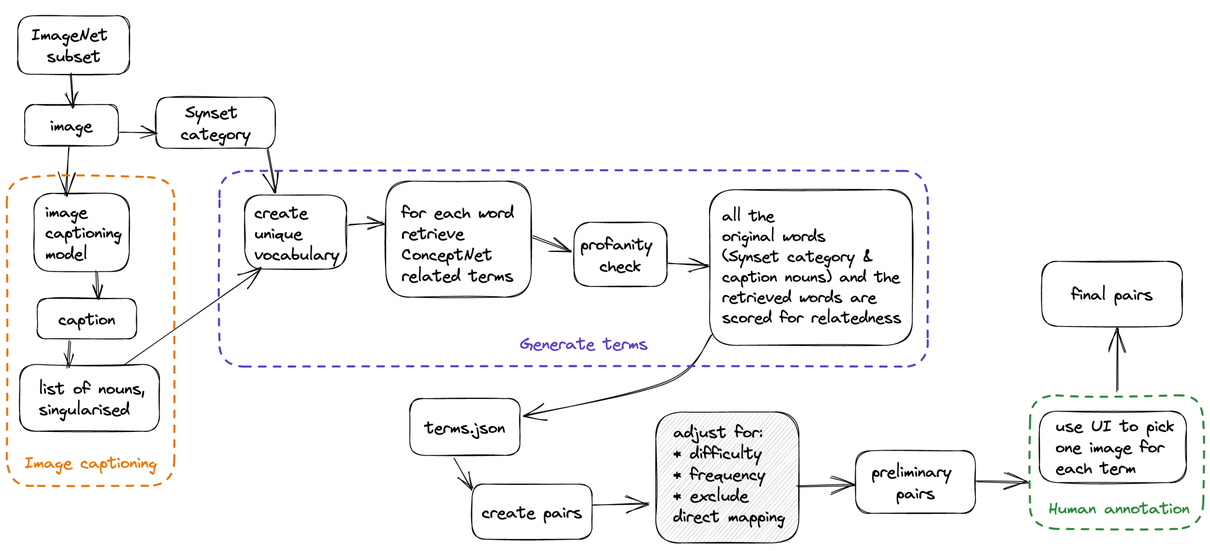


Image-Wordle

The Image-Wordle is a collaborative, image-grounded, word-guessing game. Two participants play together to guess the word explicitly or implicitly found in the image that serves as a prompt. The participants must agree on the word and enter it simultaneously. They can use the chat to coordinate the guess and discuss the image and its concepts. Letter highlighting assists the process by marking the correct letters in the right position with green and the correct letters in the wrong position with yellow.



Methodology

The image on the left visualises the pipeline from an image dataset to image-term pairs. AMT data collection setup details:

- **3 words** per session between two players
- **6 tries** per word with letter highlighting
- 1st word "easy", 2nd "just right", 3rd "difficult"
- Both or just one person sees an image(image hops)
- Corpora **frequency** setting: **3+** (at least 1 occurrence per million words)
- **Difficulty**/relatedness setting: **easy = 1**, **just right = 0.99 – 0.69**, **difficult = <0.69**

EXTERNAL RESOURCES

- **Images**: ImageNet. Subset: ILSVRC 2012, validation
- **Terms**: ConceptNet API, related terms
- **Corpora frequency**: wordfreq (Python package)
- **Profanity check**: profanity-check (Python package)
- **Singularising nouns**: inflect (Python package)
- **Image captioning model**: nlpconnect / vit-gpt2-image-captioning (Huggingface)

Preliminary findings

The preliminary data shows that the human-annotated ConceptNet relatedness score is promising in modelling the mental hops required to get from image to term, the smaller values symbolising multiple mental hops. Further, the dialogues collected reflect the assumed difficulty: the easy pairs not eliciting the desired type of dialogue and the more difficult pairs timing out without resolution.

EXAMPLE DIALOGUE

- A: "A scubadiver"
- A: "Looking at other scubadiver"
- B": we can start with DIVER"
- B: "is there water in it ?"
- A": "One is in the water and other is looking at him through a glass pane just like a fish in the acquarium"
- A: "Diver"
- *A & B enter "diver" as a guess*
- B": "next try GLASS maybe"
- A" "Or it's like a aspiring diver looking at an image of another diver"
- B: "Ok Glass it is"
- *A & B enter "glass" as a guess*
- A: "Let's try scuba"
- B": "yep"
- *A & B enter "scuba" as a guess* (correct)