Weak Supervision for Grounding Preconditions in Images

Amani Maina-Kilaas (HMC), Ehsan Qasemi (ISI), Muhao Chen (ISI)

If interested, contact amainakilaas@hmc.edu

This work was performed under the USC ISI REU Site program supported by NSF grant #2051101

Introduction

Commonsense reasoning is on the frontier of natural language understanding!

Here, we focus on learning the **preconditions** of general **statements**.

The dog has an owner / **The dog is wild** — ENABLES / DISABLES — **The dog is a pet** / **The dog is a pet**

precondition — *statement*

Contribution

We provide a visio-linguistic **dataset for multimodal models**, consisting of precondition-statement pairs along with images to help contextualize their relationship.

the traffic is stopping
ENABLES
the biker can cross

the person scolded by professor
ENABLES
the person keeps their mouth shut

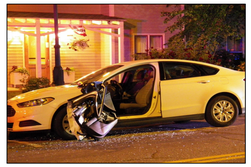
the store has closed
DISABLES
the person goes to the store one day

the person wants to learn how to drive
ENABLES
the person buys a car

Approaches

Extraction from Captions

Utilizing common linguistic patterns, we extract new statements and preconditions from existing captions.



Caption:
I would love to go for a ride in my new car, **except** someone destroyed the door last night!

Statement:
I would love to go for a ride in my new car

Precondition:
Someone destroyed the door last night!

Precondition-Caption Similarity

Using existing preconditions, we find images that have been labeled with similar captions.

Statement:
Apples are edible

Precondition:
The apple is rotten

Caption:
My new rotten apple art :)

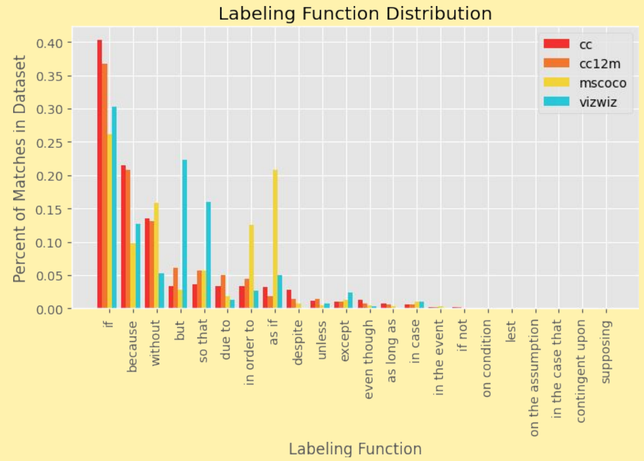
Image Querying

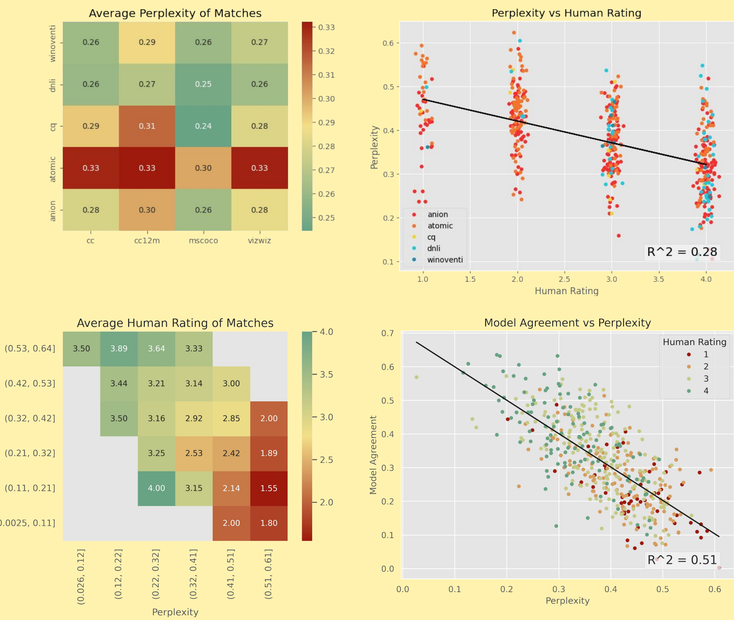
Using existing preconditions, we query directly for images using tools such as search engines.

Statement:
It is hot in the summer

Precondition:
You are in Antarctica

Dataset Statistics





Next Steps

- Refine approaches to generate cleaner data.
- Evaluate the quality and utility of our dataset.