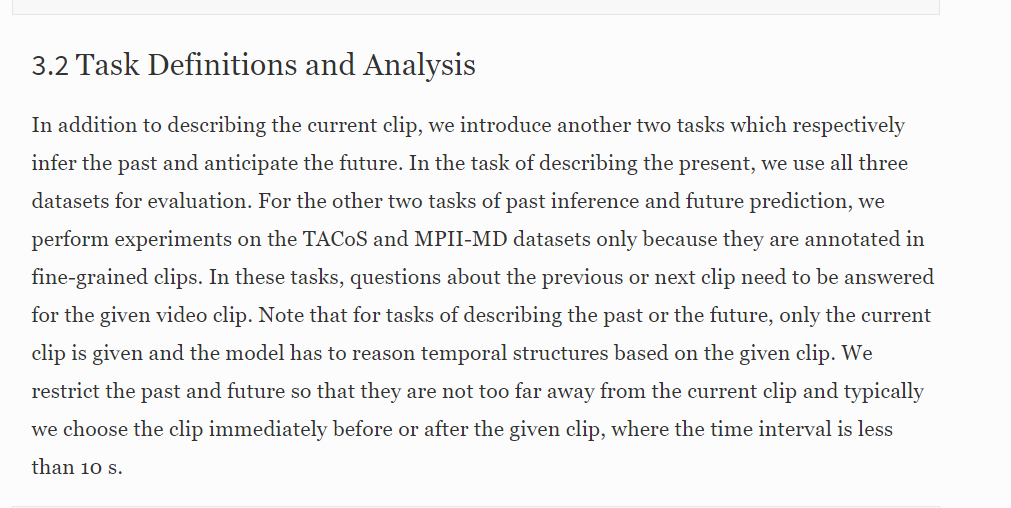
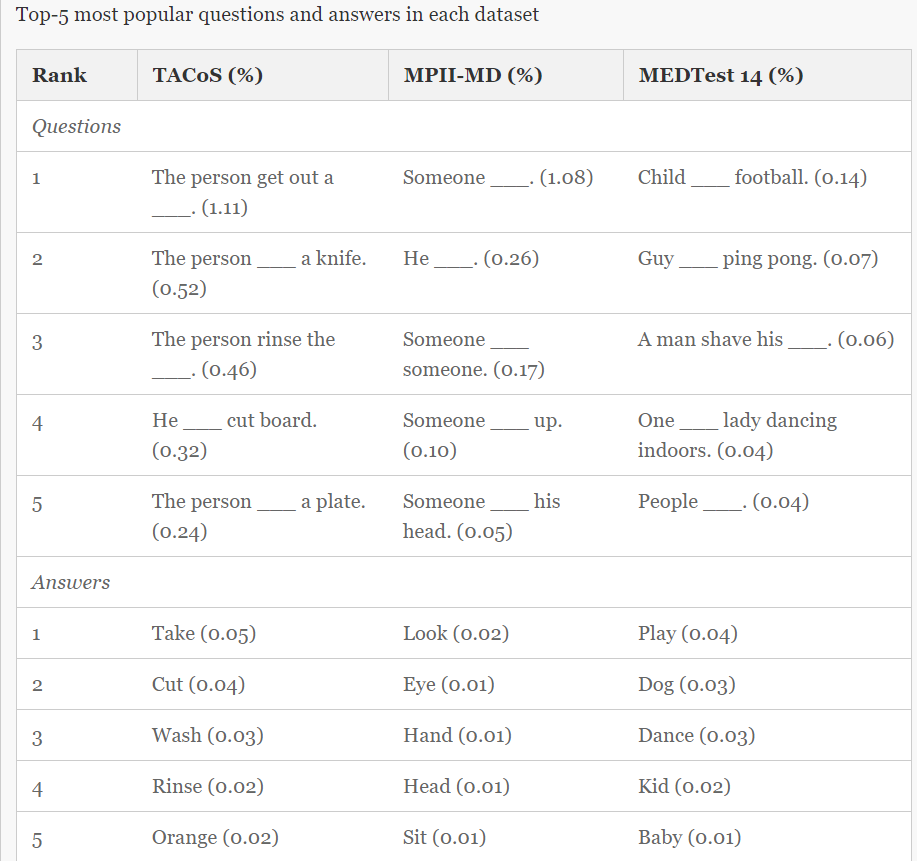
**Notes on Papers on Video Question Answering**

**MovieQA Paper** (2016 Paper):

This paper has data of many movie clips, along with their summaries from wikipedia, imd websites, etc. The goal is to get summary and answer specific questions about the plot of films.  
This paper introduced the dataset mainly.  
**MovieQA has multiple choice answers.**

**Uncovering Temporal Context for Video Question Answering** Linchao Zhu, Yi Yang….:

This paper seems more relevant than previous (**perhaps most relevant**)  
  
Dataset Details  
  
**Types of Questions**  
A) Nouns  
B) Verbs /Actions  
C) Phrases  
  
What I observed was that the videos may be short and restricted to one type of scene (not too much of movement)  
So, the question seems to be for video clip as whole.  
  
  
  
This is the section where task is defined.  
There are three tasks:  
1) Describe the current clip  
2) Predict on future clip  
3) Predict past clip  
  
There are easy and hard types of questions.  
**The questions here are fill in the blank type of questions  
They are multiple choice answer type questions, with options given.**Of these multiple answers, some of them are distractors / hard negatives.  


Word Representations: To read.  
Section 4 is the most important, which talks about the model and representation.  
**How is loss computed when multiple choice answers?**  
Dual Rank loss.  
 *Limitations:*1. the videos may be short and restricted to one type of scene (not too much of movement)

2. Multiple choice answers, so training with candidate answers, and so training end to end learning. Ideally can’t work like that.

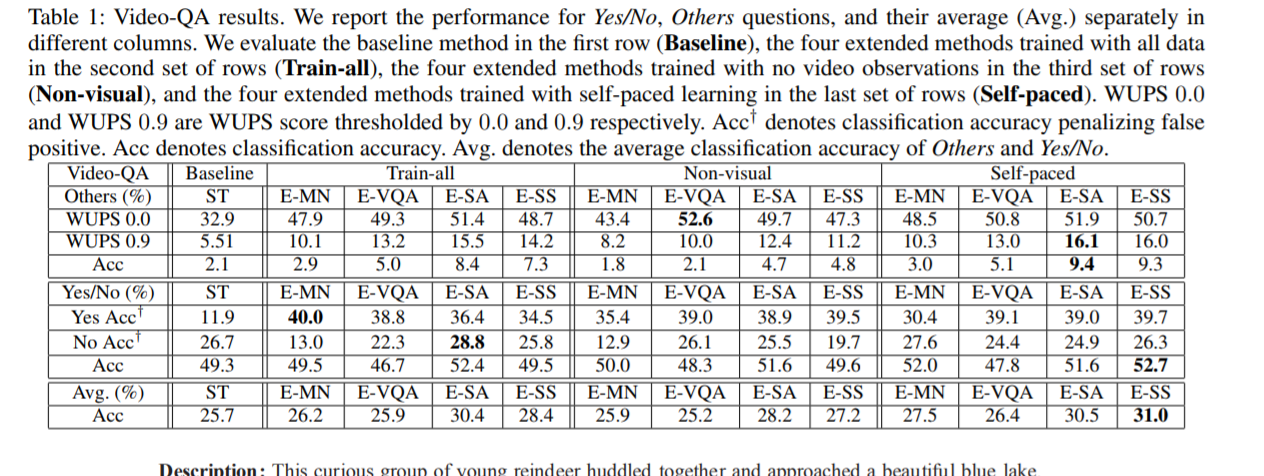
3. Fill in the blank type questions, about noun / object, action / verb and phrase.

**Leveraging Video Descriptions to learn Video Question Answering**

Dataset: (<http://jukinmedia.com/videos>)

**Strengths**: It doesn’t employ multiple choice answers like previous 2 papers.  
 Free form questions and answers.

**Weakness** *The question and answer generation has been done from Video descriptions.*

This is the major weakness which I found. The video descriptions would be very specific to the particular video, and may not encompass all the information about which question could be asked. If proper evaluation is to be done, then we need similar video with opposite things happening, for seeing if action is being recognized.  
The evidence can be seen from the results.  
  


This is the table showing accuracy. The Non-visual section / column means video hasn’t been used for coming up with answer. That column itself shows comparable accuracy.  
This may be because the features from the descriptions seem to be given importance.  
If this method is to work, we may need very huge data encompassing all types of actions and objects.

**Video Question Answering via Hierarchical Spatio Temporal Attention Networks** (Zhou Zhao, Qifan Yang)  
This may also be **very relevant** paper, as it describes some of the steps I have been thinking of.

**Weakness:**  
Here, again Question Answer pairs are generated from video descriptions.  
But the question types are only **Object**, **Number**, **Color** and **Location**There doesn’t seem to be questions on actions.  
  
**Strength:**The idea of finding out the relevant frames as the first step in pipeline seems right, like how I had thought.

**TGIF-QA Paper (CVPR)**

**Strengths:**It has additional contributions in terms of types of questions, which are:  
A) Repetition counts (example: How many times cat licks)  
B) State transition (example: What does cat do before standing up)  
  
**Weakness:**The videos are GIFs, which generally have whole frame in content, with very slight modifications throughout the GIF length.  
Eg: If a question is how many times is the cat licking itself, mostly throughout the GIF the cat would be visible and in the frame.  
Translating to longer videos, we may not know how it would consider set of frames to be relevant.

**Our Problem:**

We are working on kitchens datasets. So we are going to have lots of interactions with different kinds of objects. Primarily, the action movements by the hands must be recognized, along with the objects the hands are interacting with.