Guidelines for image description error analysis

Anonymous

April 13, 2017

1 Introduction

This document provides guidelines for the annotation of automatically generated image descriptions. Our goal is to assess the semantic competence of image description models. In other words: are the descriptions at least 'technically' correct? This is a low bar, as we ignore fluency and usefulness, which are also desirable properties for an NLG system. We define two tasks:

1. **A binary decision task**, where annotators judge whether or not a description is congruent with an image.
2. **A categorization task**, where annotators select error categories that apply for incongruent descriptions.

These tasks are strongly related: if a description is incongruent, it should fall into one of the error categories, and vice versa. Hence, annotators for either task need to be familiar with our taxonomy of errors.

2 Error categories

All our error categories are provided in Table 1. There are four main categories: People, Subject, Object, and General. I tried to strike a balance between specificity and amount of categories. No doubt some of these could be further subcategorized, but more categories means the annotation task might become overwhelming.

People	Subject	Object	General	General
Age	Wrong	Wrong	Stance	Scene/event/location
Gender	Similar	Similar	Activity	Other
Type of clothing	Inexistent	Inexistent	Position	Color
Color of clothing	Extra subject	Extra object	Number	Generally unrelated

Table 1: Error categories for incongruent image descriptions. The organization of these categories corresponds to the organization of the categories in the annotation environment.

2.1 Short description

Here's a short description of each category, and each of the subcategories. The next section provides examples for each of these.

People Image description models often make mistakes that are specific to the description of people. Subcategories are AGE (e.g. *woman* instead of *girl*), GENDER (*man* instead of *woman*), TYPE OF CLOTHING (*shirt* instead of *jacket*), and COLOR OF CLOTHING (*red shirt* instead of *blue shirt*).

Subject Mistakes relating to the subject of the description. We use the following subcategories: WRONG when the wrong entity in the image is chosen as the subject, SIMILAR when the image description system mis-identifies the subject for something visually similar (e.g. *guitar* instead of *violin*), INEXISTENT when nothing close to the mentioned entity is present in the image, and EXTRA SUBJECT/OBJECT when an additional (nonexistent) entity is mentioned besides the correct entity.

Object See **subject**.

General Mistakes that are not specific to people. The subcategories are as follows: STANCE for posture-related mistakes, ACTIVITY for wrongly identified activities, POSITION for mistakes in spatial relations within the image, NUMBER for any counting errors (too few/many entities mentioned), SCENE/EVENT/LOCATION for misidentifications of the scene, event, or location, COLOR for non-clothing entities that are mistakenly said to have a particular color, OTHER for any unforeseen mistakes, and GENERALLY UNRELATED for generally unrelated descriptions, that are beyond repair. This is usually the case when more than 2–3 error (sub)categories are applicable.

2.2 Examples

A **man** is climbing a rock
Category: Age

A man in a **blue** shirt and blue jeans is working on a ladder
Category: Color of clothing

A **girl** playing soccer
Category: Gender

A **boy** jumps over a hurdle
Category: Wrong subject

A girl in a yellow **shirt** is standing on the beach
Category: Type of clothing

A woman in a blue shirt is standing in front of a blue car
Category: Inexistent subject

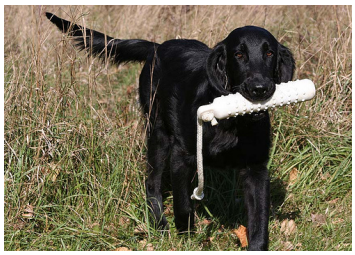
Two police officers are posing for a picture
Category: Similar subject, number

A man with a tennis racket **and a tennis racket**
Category: Extra object

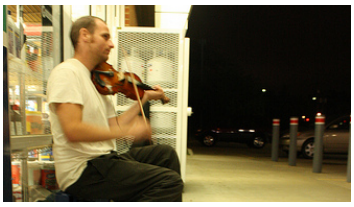
A man in a white shirt **and a man in a white shirt** are preparing food
Category: Extra subject

A man in a brown jacket is **standing** in front of a wall
Category: Stance

A young boy is holding a **little girl**
Category: Wrong object

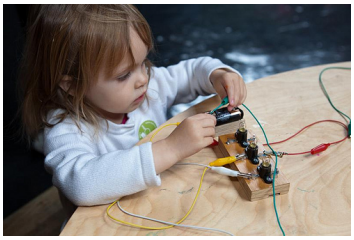


A black dog **runs through** the grass
Category: Activity



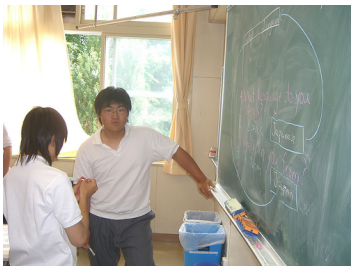
A man is playing a **guitar**
Category: Similar object

Two men are playing instruments
Category: Number



A young girl in a white shirt is playing with a **guitar**
Category: Inexistent object

A little girl in a white dress is walking **in** the water
Category: Position



A man in a white shirt and a woman in a white shirt are standing **in a hallway**
Category: Scene/event/location

A black **and white** dog is playing in the snow
Category: Color

A group of people standing in the snow
Category: Generally unrelated

A group of people are standing in **a fire**
Category: Other

2.3 Important contrasts

While the categories are fairly straightforward, there are cases where it is easy to get confused between a pair of categories. Here are additional guidelines for difficult cases that I have encountered.

- STANCE versus ACTIVITY: Use the former when the difference is static, e.g. *standing* vs. *sitting*. Use the latter if the difference is dynamic, e.g. *standing* versus *walking*.
- SCENE/EVENT/LOCATION versus POSITION: Use the former when the surroundings are not correct. Use the latter when position within the surroundings is not correct.
- EXTRA SUBJECT/OBJECT versus NUMBER: Use the former when the subject/object is wrongfully extended with a conjunction (e.g. *and a woman in a white shirt*). Use the latter when there's a general mismatch in number (*a, one, two, three, a group of*).
- SIMILAR OBJECT versus POSITION: This conflict arises in cases where e.g. *. . . is sitting on a bench* is used instead of *. . . is sitting on a chair*. In all these cases, use *similar object*. (Even if there is an actual bench in the image.)

3 Task descriptions & instructions

Now that we have seen the different error categories, we can describe the two main tasks as follows:

Task 1: Congruency Judge whether the generated description is congruent (no error categories apply) or incongruent (at least one error category applies).

Task 2: Categorizing incongruent descriptions Annotate the 'semantic edit distance' between the generated description and the closest valid description that you can imagine. Tick all the error categories corresponding to the things you would have to change. If the generated description is unrelated to the image, or if you feel that there are too many changes necessary to get to a valid description, select GENERALLY UNRELATED.

The threshold for when a description is generally unrelated is undefined. In general, I feel like type/color of clothing don't really hurt the relation between description and image as much as e.g. having the wrong verb. So it all comes down to your intuition.

4 Evaluation: correcting the errors

This is a separate task that serves both as an evaluation of Task 2, and as an indication of system performance if all errors identified in Task 2 are addressed. The correction task works as follows.

1. Select an error type to correct. E.g. COLOR OF CLOTHING.
2. Go through all images annotated with this type, and correct *only* the relevant error.
3. When all relevant errors are corrected, we evaluate the results using BLEU/Meteor.

It is important for this task to be conservative in editing the descriptions. Try to change as little as possible. If a change would require restructuring the entire sentence, leave the description as it is. We'd rather underestimate than overestimate the improvement from fixing the errors. Otherwise we'd just be evaluating how good humans are at writing descriptions. So e.g. for colors, *only* change color terms into other color terms. For gender, only change *man* $\leftrightarrow$ *woman* and *boy* $\leftrightarrow$ *girl*, not *man* $\leftrightarrow$ *girl*. That would be changing the age along with the gender.