Pragmatic descriptions of perceptual stimuli

Emiel van Miltenburg
Vrije Universiteit Amsterdam
emiel.van.miltenburg@vu.nl

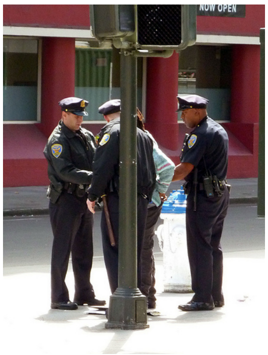
Abstract

This research proposal discusses pragmatic factors in image description, arguing that current automatic image description systems do not take these factors into account. I present a general model of the human image description process, and propose to study this process using corpus analysis, experiments, and computational modeling. This will lead to a better characterization of human image description behavior, providing a road map for future research in automatic image description, and the automatic description of perceptual stimuli in general.

1 Introduction

Automatic image description is a key challenge at the intersection of Computer Vision (CV) and Natural Language Processing (NLP), because it requires a deep understanding of both images and natural language (Bernardi et al., 2016). There are two major datasets that are used to train and evaluate automatic image description models: Flickr30K (Young et al. (2014); 30K images) and MS COCO (Lin et al. (2014); 150K images). These descriptions were collected through a crowdsourcing task where Workers were asked to provide one-sentence descriptions for each image. One of the assumptions behind these datasets is that they provide objective image descriptions:

> "By asking people to describe the people, objects, scenes and activities that are shown in a picture without giving them any further information about the context in which the picture was taken, we were able to obtain conceptual descriptions that focus only on the information that can be obtained from the image alone." (Hodosh et al., 2013, p. 859)



Human: Three policemen are standing around someone in a gray sweatshirt with stripes.

Model: A group of people are walking down the street.

Figure 1: Flickr30K image (4944749423) with a human- and a machine-generated description.

The **assumption of neutrality** is a useful simplification: if it is more or less correct that similar images will have similar descriptions (that are not influenced by any external factors), then we can try to learn a mapping between images and descriptions. This is what Vinyals et al. (2015) do. They use a Long Short-Term Memory model to generate sequences of words, given the visual context.[1] Their model is able to produce reasonably good image descriptions without using any higher-order reasoning. Figure 1 provides an example.[2] The machine-generated descriptions are typically shorter and more general than human descriptions. For example, the model talks about 'a group of people', rather than about *a group of policemen* and *a civilian*. Compared to humans, there is less variation in the kind of labels that the model uses to refer to people (section 2.5 of this

[1] The visual context was provided by a convolutional neural network model (Ioffe and Szegedy, 2015), trained for the 2014 ImageNet Large Scale Visual Recognition Challenge (Russakovsky et al., 2015).

[2] More examples at https://github.com/evanmiltenburg/NIC-data.

Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 1–10, Valencia, Spain, April 3-7 2017. ©2017 Association for Computational Linguistics

paper). And for a good reason: human-level specificity requires a deeper understanding of context.

This proposal challenges the neutrality assumption, and aims to characterize the subjective nature of image descriptions. Such a characterization is necessary to get an overview of the challenges that lie ahead. My main thesis is that *image description is not a simple mapping from visual features to strings of words. Rather, it is a process involving reasoning, perspective and world knowledge*. This thesis is supported by empirical evidence from image description corpora, showing how the descriptions reflect the crowd-workers' *interpretation* of the images. I will investigate what are the limits of current image description systems, and what is needed in order to get human-like performance, using the model in Figure 2.

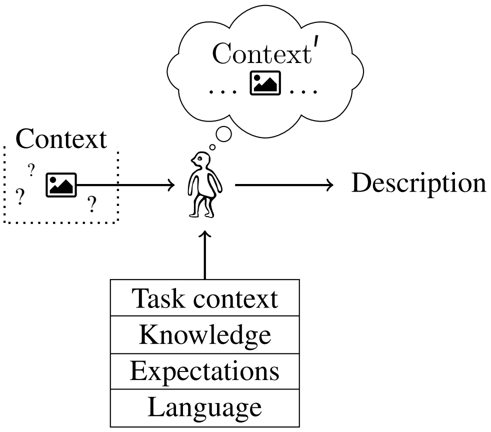


Figure 2: Conceptual model of description generation. Note that the original context is likely to be different from the context inferred by the subject.

In this conceptual model (corresponding to the data collection procedure for Flickr30K and MS COCO), an image is taken out of its original context and presented to a human annotator. Because the original context is lost, the annotator now has to *re-interpret* the image within the context of the task. This new understanding is based on their world knowledge and prior expectations. Next, the annotator has to verbalize his or her understanding of the image in one sentence. This means they have to make choices about (1) which aspects of the picture to focus on, and (2) the way in which those aspects should be described. The first is limited by relevance, whereas the second is limited by the linguistic means afforded by the annotator's native language.

The main goal of this proposal is to better understand the process of describing an image, from the speaker's point of view. In other words: how does someone 'come up' with a description for a given image? How do they determine which features to include? I will use a three-pronged approach to answer these questions:

Corpus analysis Looking for patterns in large volumes of uncontrolled image description data. Our main goal here is to characterize and quantify image description behavior. I will discuss my approach in §2.

Experiment Studying phenomena discovered through corpus analysis in a controlled setting. Our main goal here is to understand what factors drive these phenomena. Corpus analysis and experimentation are two sides of the same coin: we can use a corpus to generate hypotheses about how people describe images, and use experiments to test those hypotheses. See section 3 for more.

Modeling Studying the capacity of automatic image description systems to capture pragmatics. Our main goal here is to characterize the gap between human and machine performance; what makes image description difficult, and how could we face those challenges? Discussed in §4.

The result of this approach is an interdisciplinary picture of the image description process, combining insights from linguistics, natural language processing, and social science.[3] Because of the social relevance of image description systems (one of the main motivations to build these systems is to make images accessible for the visually impaired), I will also discuss the ethical implications of this research (§5). Finally, section 6 takes the first step in generalizing our model to other modalities. Specifically, I discuss similarities and differences between visual and auditory stimuli.

2 Corpus analysis

I will analyze several different image description incorporate in order to characterize the way that people talk about images. The study is not limited to English data sets, but also extends to other languages, allowing us to explore the differences between speakers of different languages. This section provides an overview of the work I have already done (van Miltenburg, 2016; van Miltenburg et al., 2016a, in §2.1–2.4), as well as some work in progress (§2.5,2.6).

[3] Fully in the spirit of Krahmer (2010), who argues that computational linguists can learn a great deal from psychologists (and vice versa).

2.1 What to include in a description

Table 1 shows the ways in which a phrase or expression can be related to an image.[4] It can either refer to something inside or external to the image, and annotators can choose whether or not to use it in their description. This choice is called **content determination/selection** in Natural Language Generation (Reiter and Dale, 1997; Reiter and Dale, 2000). I propose to use this table to systematically study how humans perform this task.

	In the image	**Image-external**
In the description	A	B
Not in the description	C	D

Table 1: Ways in which an expression or a phrase can be related to an image.

The label *policemen* in Figure 1 is an example of **situation A**: there are policemen in the image, and the annotator decided to include that expression in their description. Two other annotators in the Flickr30K dataset also speculated about *an arrest* taking place:

(1) Other descriptions of the image in Figure 1.

a. Three officers arresting someone on the corner of a street.

b. Police officers are arresting a woman.

These are examples of **situation B**, because we cannot conclude this from the image alone. Maybe the person in the gray sweatshirt had just fallen and the officers are helping them stand up. In (van Miltenburg, 2016), I call cases like this **unwarranted inferences**, and provide a list of different kinds of these inferences in the Flickr30K dataset. For example, RELATION-inferences where young children are assumed to be siblings or friends, or women with children are assumed to be mothers.

Traffic light is a nice example of **situation C**: there is a big metal pole right in front of the policemen, but the annotator in Figure 1 made no reference to it. They were also careful not to say that the policemen are *arresting someone*, even though two other annotators did make that inference. This is an example of **situation D**. Note that we are only able to identify this situation because the other annotators *did* speculate about the situation in the image.

2.2 Marking

Following Jakobson (1972) and others in linguistics, we will use the term **marking** to denote the act of signaling an entity or attribute. The difference between Situation A and C is one of *markedness*. We can ask ourselves *why* annotators decide to mark some entities or attributes, but not others. The most basic and naive explanation is that this is because *those are the most important*. But this only gets us part of the way. There is a large amount of variation in the entities and attributes marked by the different annotators in the Flickr30K and MS COCO corpora. An additional explanation is grounded in the work by Beukeboom (2014), who argues that the kind of language people use reflects the way they view the world (he calls this **linguistic bias**); since the annotators' perspectives differ, so do the descriptions.

2.3 What gets marked?

People typically mark entities, properties, or events that are unexpected or go against some social norm (Beukeboom, 2014). Negations are the clearest example of this. Example (2) shows two descriptions from the Flickr30K corpus, where annotators explicitly marked what the subjects in the images *weren't* doing, so as to emphasize that this behavior is unusual.

(2) Examples from van Miltenburg et al. (2016a)

a. Man **not wearing a shirt** playing tennis.
↰ You are supposed to wear a shirt.

b. A boy is eating pie **without utensils**.
↰ You are expected to eat with utensils.

At the same time, there are also structural differences between *a priori* comparable entities or groups of entities in the way they are marked. I annotated all pictures of babies in the Flickr30K dataset, and found that 22% of all black babies are marked as 'black' or 'African-American', and 14% of all asian babies were marked as 'asian' or 'oriental', while less than 1% of all white babies are marked as such (van Miltenburg, 2016). For the group of Flickr30K annotators, it seems that 'white' is the expected default and thus less marked than the others.[5]

2.4 Negations, norms and expectations

As mentioned above, the Flickr30K data contains several descriptions containing negations. The

[4] For a detailed taxonomy of the inverse relation —ways in which an image can relate to a text, see (Marsh and Domas White, 2003).

[5] This is related to reporting bias, see (Misra et al., 2016).

examples in (2) are surprising: somehow crowd workers decided that the best way to describe the relevant images is to say what is missing from them. This behavior is a result of our everyday experience, which (along with social norms) gives rise to expectations about how people are supposed to behave. Negations provide a linguistic means to signal mismatches between our expectations and what is actually happening (Beukeboom et al., 2010). Now consider the items in (3):

(3) a. not wearing a shirt (negation)
b. wearing a blue shirt (specification)
c. wearing a shirt (unmarked)

Negations like (3a) not only signal deviations from the norm, they also (indirectly) tell us *what the norm is*. The same can be said for modifiers further specifying a noun phrase, e.g. (3b); if there are examples of further specification of a noun phrase, but no examples of the 'plain' noun phrase, then we know that the noun phrase corresponds to the norm. Examples like (3c) are typically only used to signal a contrast (in this case: with others not wearing a shirt). I will try to use observations like these to find out what are the implicit norms in image description datasets.

van Miltenburg et al. (2016a) provide a categorization of the different uses of negation, in order to gauge the kind of background knowledge that is required to produce descriptions containing negations. These range from signaling that someone is not wearing a shirt or not using utensils (2) to image-specific cases like (4):

(4) Several people sitting in front of a building taking pictures of a landmark **not** seen.

Here, a crowdworker concluded that the people in the image must be taking pictures *of a landmark*, without having seen the actual landmark. Negations in the Flickr30K data signal cases where world knowledge and commonsense reasoning is required for generating descriptions. This makes descriptions with negations a suitable paradigm to evaluate the extent to which automatic image description systems are able to generate humanlike output. I will check whether state-of-the-art image description systems are able to produce negations and, if so, what kind of negations they are able to produce (see also §4). My expectation is that the production of negations will be limited to common cases, where entire phrases containing negations can be reproduced from the training data. Going beyond those cases requires higher-level reasoning, which current models are not designed to perform.

2.5 Labeling

What kind of labels should be used to refer to people? Figure 3 (next page) shows that there is a large variety of labels in the Flickr30K dataset, that belong to different semantic categories.

Occupation police officer, businessman, shepherd.
Relation grandma, boyfriend, colleague, neighbor.
Activity speaker, activist, presenter. Subcategories:
  Sports snowboarder, athlete, football player
  Music trumpet player, saxophonist, pianist
Age toddler, boy, girl, adolescent, adult.
Gender male, female, boy, girl, man, woman.
Appearance redhead, blonde.
Religion hindu, muslim, jew.
Other vagabond, nerd, idiot.

Figure 3: Kinds of person-labels in the Flickr30K dataset, with examples. Subcategories are dominant, coherent subsets of the data.

When annotators decide on a label to use, they can roughly base their judgment on two factors: **appearance** and **situation** of the person to be labeled. Table 2 provides a categorization of person labels in terms of these factors. In the data collection process, I noticed that it was quite easy to find examples of mostly appearance-based or mostly situation-based labels, but difficult to find good examples of labels that seem to depend equally on both appearance and situation. *Civilian* is a good example, because felicitous use of this label requires the relevant person to be around e.g. military personnel (the situation) while not wearing a uniform themselves (appearance).

We can also think of the labels in Table 2 as being on a continuous scale showing the reliance on either of these two factors, as shown in Figure 4. To be clear: I do not want to claim that the use of 'civilian' is somehow *less situation-based* than the use of 'neighbor'. Rather, it balances between two forces that drive the labeling process.

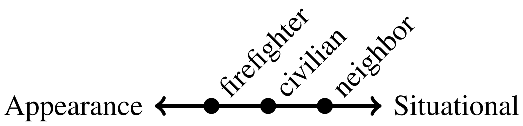


Figure 4: Continuous scale from Appearance-based to Contextually determined labels.

I will further formalize the taxonomy from Figure 3, and extend it to include adjectives and other

Appearance	Situation	Example
Yes	No	Police officer, businessman, firefighter
Yes	Yes	Civilian
No	Yes	Bystander, neighbor, passerby, orphan
No	No	—

Table 2: A categorization of labels based on whether the label is applied on the basis of someone's appearance or the situation they are in.

modifiers, as well as mark each category for its reliance on appearance and situation. I will then study differences in the use of these labels between human annotators and automatic image description systems. We can also use this data as a guide for image description models to produce or not to produce particular kinds of labels. At the same time, this data is useful for natural language understanding as well: with a resource telling us what alternatives a speaker may have in referring to a particular entity, we can reason over *why* the speaker said X while they could have also said Y or Z (Grice, 1975; Geurts, 2010).

2.6 Cross-linguistic analysis

There is a growing interest in collecting image descriptions in different languages, so as to be able to generate descriptions in languages other than English (e.g. Chinese (Li et al., 2016), German (Elliott et al., 2016), Turkish (Unal et al., 2016)). This enables us to study how speakers of different languages describe the same images. Some examples of differences between languages can already be found in the literature. For example: Li et al. (2016) provide the example of an image with a woman taking a picture. In the English descriptions, the woman is referred to as *an Asian woman*, whereas in the Chinese descriptions she is described as a *middle-aged woman* (presumably because *Asian* isn't a distinctive feature in China). Later, the authors note about the English descriptions translated to Chinese that they "do not necessarily reflect how a Chinese describes the same image." This is in line with our model (in Figure 2), which shows the influence of knowledge, expectations, and language on the image description process.

I will study the influence of language by collecting Dutch image description data, and comparing this data with English and German descriptions, so as to see whether the Dutch crowd workers display any behavior that is different from the German and English workers. For example, whether Dutch annotators use different kinds of labels than German or English annotators. The reason for collecting Dutch descriptions is that our project is based in the Netherlands, and if we discover any interesting phenomena, we will be able to carry out additional lab experiments with Dutch participants to further explore those phenomena.

3 Experiment

What makes the crowd describe images the way they do in the MS COCO and the Flickr30K data? I will investigate the degree to which the format of the crowdsourcing task affects the descriptions, and how we can get people to provide different kinds of descriptions. Experiments are essential to test hypotheses that arise from the analysis of image description corpora. I will discuss two experiments below (sections 3.2 and 3.3), but first let us look at the format of image description tasks.

3.1 Canonical format

I will refer to the Flickr30K and MS COCO annotation tasks as the *canonical format*. In this setup, a task consists of a set of general instructions and examples of 'good' and 'bad' descriptions, followed by a set of five images with a prompt to describe each image in one complete, but simple sentence. Crucially, the participants are not told why they are providing the descriptions, or how the descriptions will be used.

3.2 Speculation

Even though the instructions explicitly tell workers not to speculate, we can find many cases of unwarranted inferences. This seems to go against Grice's (1975) Maxim of Quality ("try to make your contribution one that is true"). Assuming that Workers do try to be helpful, my hypothesis is that this behavior is a direct result of the canonical format: left wondering how their description will be used, Workers just provide as much information as possible because the *question under discussion* is unclear (Roberts, 1996). I plan to test this hypothesis by changing the prompt (specifying how the descriptions will be used) and collecting new descriptions for a subset of the images in the Flickr30K dataset. I expect that the new prompt will make the elicited descriptions more concise and uniform, because participants will focus more on the central aspects of the images that are relevant to the proposed application.

3.3 Entrainment and differentiation

Entrainment and differentiation are well-known effects where speakers either keep re-using the same phrase to refer to the same or similar entities, or change their phrasing to contrast new entities with others (van der Wege, 2009). These within-subject effects have been mostly been studied in the lab with small amounts of abstract examples, and I will use crowdsourcing to extend this research to photographs on a large scale.

To find out whether there are such within-subject effects in the MS COCO and Flickr30K data, it is necessary to know who provided which description, and in what order the images were presented. Because the raw crowdsourcing data with Worker IDs has not been released for the Flickr30K and MS COCO data, we do not know the extent of these effects in image description data. I have contacted the authors to obtain the raw data, and also plan to set up a controlled study to measure entrainment and differentiation effects.

In this study, I will present sets of images in different orders, and collect a large amount of descriptions for each ordering. After collecting this data, I will analyze the data for entrainment or differentiation patterns. This work can also be seen as a more general test of the assumption that the image descriptions in MS COCO and Flickr30K are *independent* from each other. If it turns out that the other images in the task influence the way an image is described, then this effect needs to be taken into account.[6] At the same time, entrainment and differentiation effects are very informative about how people deal with similarity and differences between images, and we should try to see how these effects can be leveraged to create better performing image description systems.

3.4 Related work: Stylistic variation

There is already some prior work showing that the way that crowd workers are prompted for a description can have a strong influence on form of the descriptions. Baltaretu and Castro Ferreira (2016) present results from a study manipulating a crowdsourcing task to get different kinds of referential expressions for the same entity. They experimented with different task prompts within the ReferIt-game (Kazemzadeh et al., 2014). In this annotation game, participants are asked to provide referring expressions for specified entities. They score points if other participants can successfully identify the entity from the referring expressions. Baltaretu and Castro Ferreira (2016) modified the original prompt by asking participants to play fast (FA), be creative (CR), be clear and thorough (CT), or just to provide descriptions without any additional goal (NO). These different prompts had an effect on the length of the expressions (with longer expressions in the CR and CT conditions), and on the amount of adjectives used (with more adjectives in the CR-condition than in the FA-condition). Table 3 shows an example description for each category.

FA	Jumping monkey.
CR	A primate showing off his business end.
CT	Small monkey with a very long tail.
NO	A monkey on a person’s head.

Table 3: Example from Baltaretu and Castro Ferreira (2016), showing the difference between the different prompts: Fast, Creative, clear and thorough, and no specific emphasis.

An important observation is that human language is capable of enormous variation. The richness of language poses many challenges to developers of image description systems. For example: when do you use what kind of description?

4 Modeling

Models are essential to our understanding of the world. By building a system that is able to describe an image exactly as a human would do, we can demonstrate that we understand the entire image description process. But right now, we are still far from reaching that goal. In this project, I will try to lay out a road map for the future, by looking

[6] To some extent, this is already controlled for in the current datasets, as Mechanical Turk and Crowdflower randomize crowdsourcing tasks. But this only means that the five descriptions per image are each primed in a different way.

at the discrepancies between human performance and the performance of state-of-the-art models. I plan to carry out three kinds of studies:

Evaluation and error analysis Evaluation of image description systems is typically done by running a metric comparing the generated output with a set of reference descriptions produced by human annotators (see (Kilickaya et al., 2017) for an overview). The problem with these measures is that they are very coarse-grained. I am currently working on a manual error analysis, checking whether automatically generated descriptions are fully congruent with the relevant image, or whether there are any mistakes. Annotating all the mistakes allows us to classify and then quantify which mistakes were made how many times. The error categories show us where there is still room for improvement.

Producing particular phenomena Having made several observations in image description corpora (§2), the question is whether image description systems are able to reproduce those phenomena. For example: can image description systems produce negations? (§2.4) This question calls to mind Chomsky's Competence-Performance distinction (Chomsky, 1965). When image description systems are evaluated on a particular test set, they produce one description for each of those images. This gives us a surface-level idea of their capabilities. But suppose that a system never produces a negation for any image we feed it. That does not mean that the system is incapable of producing negations. Or, putting it in cognitive terms, that it does not *know* how to use negations. It only means that negations are unlikely to be produced by the system. And so we need to dig deeper in order to find out whether the system has gained the relevant knowledge from the training data.

Generating Dutch descriptions Due to the size of the Dutch crowd, I will only be able to collect a relatively small set of Dutch image descriptions. We plan to train a machine translation system that converts English image descriptions to Dutch, so as to extend the Dutch description data. We can then train an image description model for Dutch using this extended dataset. This way we can test whether machine translation is a good strategy to develop image description systems for lesser-resourced languages. I am not the first to propose a translation-based strategy to train an image description system. Li et al. (2016) show that it's possible to train a Chinese system based on translations of the English descriptions from Flickr8K (Hodosh et al., 2013). My contribution will be to provide a qualitative analysis of the system output: does the model make different kinds of mistakes (based on the translation)? Do the descriptions sound natural?

5 Bias and ethics

As recently noted by Hovy and Spruit (2016), there has been "little discourse in the [NLP] community" about ethics, and the social impact of natural language processing. Their paper opens up the discussion, and provides some useful terminology, which can be summarized as follows:

1. Any dataset is **demographically biased**, which may lead to the **exclusion** or **misrepresentation** of social or ethnic groups.
2. Modeling data has the side-effect of **overgeneralization**.
3. "**Topic overexposure** creates biases"; useful heuristics may be disproportionately linked to particular social or ethnic groups.
4. NLP tools could be misused, or (unintentionally) further marginalize particular social or ethnic groups. These are **dual use** problems.

Some of these ideas are also discussed by Gillespie (2014), mostly in the context of information retrieval. He lists six dimensions to critically examine an algorithm, of which we will focus on the **patterns of inclusion**: how is the training data prepared, and what does it contain? We can separate two concerns for the image data under discussion: image selection and annotator selection.

Image selection Both datasets are based on images from Flickr. Gillespie (2014, p. 185) notes that this data may already be biased through users' interaction with the community (who value particular kinds of images) and Flickr's internal search algorithm (which also values particular images and tags). Moreover, images on Flickr also typically depict Western scenes (Miyazaki and Shimizu, 2016).

The Flickr30K images were sourced from six different groups (sub-communities set up around a particular kind of images, e.g. *strangers!* or *dogs in action*) on Flickr, and were manually selected "to depict a variety of scenes and situations" (Hodosh et al., 2013; Young et al., 2014). By focusing

on a small set of groups, one runs the risk of ending up in a 'photo bubble' where the kind of pictures in your dataset is determined by the interests of a small group of people.
The MS COCO images were collected by first compiling a list of object categories, and then looking for images containing those objects on Flickr (Lin et al., 2014). This object-driven approach means that image-sampling takes place at the community level, rather than the sub-community level. A downside of this approach is that it is language-based. Pictures taken by users who don't tag their images or who tag their images in a different language are not considered.

Annotator selection The descriptions of the images for both the Flickr30K and the MS COCO datasets were collected through Amazon's Mechanical Turk. For the former, only workers from the USA who passed a spelling and grammar test were allowed to provide descriptions. No other details about the demographics of the workers were collected. For the latter, Chen et al. (2015) note that their annotation task is strongly inspired by the annotation process for Flickr30K. Again, no details about the demographics of the workers were collected. This makes it very difficult to analyze the data for differences between groups in how they describe an image. We can say that only focusing on workers from the USA means that the descriptions all come from an American point of view. This leads to descriptions like the following, where the Otherness of the images is emphasized (all descriptions taken from the Flickr30K data):[7]

(5) a. This man is looking at shirts in a store where the language is not English .
b. I see people going into a yellow bus from another country , not United States .
c. A wild animal not found in America jumping through a field .

To get a sense of the population of Mechanical Turk, Huff and Tingley (2015) carried out a survey among United States workers asking about political attitudes and demographic factors. While there is a reasonably good overall balance between males (54%) and females (46%), the pool is racially skewed with nearly 75% White workers. Of course there might be selection bias in which workers opt for annotation tasks, but these post-hoc numbers are the best we can get. Now recall the finding that that black babies are more often marked as such (using adjectives like *black, African-American*) than white babies (van Miltenburg, 2016). This is consistent with the idea that people typically mark others who are diferent from themselves (mentioned in (Beukeboom, 2014)). Given this social dynamic, it seems clear that annotators should be selected with care. At the very least, it's worth recording more details about the crowd-workers so that we can study the effects of demographic characteristics on image descriptions.

Both image selection and annotator selection give rise to dual use issues. I will focus on the latter, because it hinges on a recurring theme in this proposal: subjectivity in language. If we better understand the processes that give rise to subjective descriptions, then we can also try to mitigate the effects of annotator bias. Through the proposed studies in the previous sections, I aim to raise awareness of the biases in image description data, and to produce a set of tools and resources that will spur improvement in this area. For example, the ability to detect whether or not a description is speculative might help to make systems deliver more factual descriptions.

6 Discussion: Other modalities

We can generalize the observations made about the image description process to other modalities. Distributional approaches to ground language in perceptual data have not only been proposed for images, but also for sounds (Lopopolo and van Miltenburg, 2015; Kiela and Clark, 2015) and even smells (Kiela et al., 2015). We also need to keep these other modalities in mind when we are working on image description, because comparing results for different modalities teaches us what is modality-specific and what is more generally true about the relation between language and perception. As a basis for future work, van Miltenburg et al. (2016b) carried out a crowdsourcing experiment to collect 'keywords' for 2,133 sounds from the Freesound database (Font et al., 2013). For the sounds that were harder to recognize, many participants resorted to speculate about the possible sources of the sound. Really understanding what a sound is about requires annotators to recontextualize the sound and think about likely events

[7] This also works the other way round. When Miyazaki and Shimizu (2016) asked Japanese workers to describe images from Flickr30K, "words such as 'foreign' and 'oversea' [initially were] everywhere in the descriptions" (p. 1783).

that may have caused it. The difference between sounds and images is that sounds are dynamic (and thus contain more information about actions than about entities) while images are static (and thus contain more information about entities).

7 Conclusion

In this paper I have proposed to study the image description process in terms of the model in Figure 2, using three different approaches: corpus analysis, lab experiments, and using image description models. This work will hopefully lead to a more complete characterization of the knowledge that human annotators bring to bear on image description tasks. This characterization will provide a road map to make automatic image description systems display more human-like behavior.

Acknowledgments

I would like to thank the members of the CLTL lab, and two anonymous reviewers for their valuable feedback. Many thanks to all my collaborators. This project is supervised by Piek Vossen and Desmond Elliott. All faults are my own. This research was supported by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen.

References

Adriana Baltaretu and Thiago Castro Ferreira. 2016. Task demands and individual variation in referring expressions. In *Proceedings of the 9th International Natural Language Generation conference*, pages 89–93, Edinburgh, UK, September 5-8. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Camiel J. Beukeboom, Catrin Finkenauer, and Daniël H. J. Wigboldus. 2010. The negation bias: when negations signal stereotypic expectancies. *Journal of personality and social psychology*, 99(6):978.

Camiel J. Beukeboom. 2014. Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies. In J. Laszlo, J. Forgas, and O. Vincze, editors, *Social cognition and communication*, volume 31, pages 313–330. Psychology Press. Author's pdf: `http://dare.ubvu.vu.nl/handle/1871/47698`.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5^{th} Workshop on Vision and Langauge at ACL '16.*

Frederic Font, Gerard Roma, and Xavier Serra. 2013. Freesound technical demo. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 411–412. ACM.

Bart Geurts. 2010. *Quantity implicatures*. Cambridge University Press.

Tarleton Gillespie, 2014. *Media technologies: Essays on communication, materiality, and society*, chapter The Relevance of Algorithms, pages 167–193. MIT Press.

Herbert Paul Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and semantics 3: Speech acts*, pages 44–58. Academic Press, New York.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany, August. Association for Computational Linguistics.

Connor Huff and Dustin Tingley. 2015. "who are these people?" evaluating the demographic characteristics and political preferences of mturk survey respondents. *Research & Politics*, 2(3):2053168015604648.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167.*

Roman Jakobson. 1972. Verbal communication. *Scientific American*, 227:72–80.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October. Association for Computational Linguistics.

Douwe Kiela and Stephen Clark. 2015. Multi- and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal, September. Association for Computational Linguistics.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, Beijing, China, July. Association for Computational Linguistics.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *To appear in Proceedings of EACL 2017*. Available as arXiv preprint arXiv:1612.07600.

Emiel Krahmer. 2010. What computational linguists can learn from psychologists (and vice versa). *Computational linguistics*, 36(2):285–294.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM International Conference on Multimedia Retrieval*, pages 271–275. ACM.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Alessandro Lopopolo and Emiel van Miltenburg. 2015. Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 70–75, London, UK, April. Association for Computational Linguistics.

Emily E. Marsh and Marilyn Domas White. 2003. A taxonomy of relationships between images and text. *Journal of Documentation*, 59(6):647–672.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2939, June.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790, Berlin, Germany, August. Association for Computational Linguistics.

Ehud Reiter and Robert Dale. 1997. Building applied natural language generation systems. *Natural Language Engineering*, 3(01):57–87.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Studies in Natural Language Processing. Cambridge University Press.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Mesut Erhan Unal, Begum Citamak, Semih Yagcioglu, Aykut Erdem, Erkut Erdem, Nazli Ikizler Cinbis, and Ruket Cakici. 2016. Tasviret: Görüntülerden otomatik türkçe açıklama oluşturma İçin bir denektaçı veri kümesi (tasviret: A benchmark dataset for automatic turkish description generation from images). In *IEEE Sinyal İşleme ve İletişim Uygulamaları Kurultayı (SIU 2016)*.

Mija M. van der Wege. 2009. Lexical entrainment and lexical differentiation in reference phrase choice. *Journal of Memory and Language*, 60(4):448–463.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016a. Pragmatic factors in image description: the case of negations. In *Proceedings of the 5^{th} Workshop on Vision and Language at ACL '16*.

Emiel van Miltenburg, Benjamin Timmermans, and Lora Aroyo. 2016b. The vu sound corpus: Adding more fine-grained annotations to the freesound database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia. European Language Resources Association (ELRA).

Emiel van Miltenburg. 2016. Stereotyping and bias in the flickr30k dataset. In Jens Edlund, Dirk Heylen, and Patrizia Paggio, editors, *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*, pages 1–4.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.