Talking about other people: an endless range of possibilities

Emiel van Miltenburg
Tilburg University
C.W.J.vanMiltenburg@uvt.nl

Desmond Elliott
University of Copenhagen
de@di.ku.dk

Piek Vossen
Vrije Universiteit Amsterdam
piek.vossen@vu.nl

Abstract

Image description datasets, such as Flickr30K and MS COCO, show a high degree of variation in the ways that crowd-workers talk about the world. Although this gives us a rich and diverse collection of data to work with, it also introduces uncertainty about how the world should be described. This paper shows the extent of this uncertainty in the PEOPLE domain. We present a taxonomy of different ways to talk about other people. This taxonomy serves as a reference point to think about how other people should be described, and can be used to classify and compute statistics about labels applied to people.

1 Introduction

There are currently two major data sets used to train and evaluate automatic description systems: Flickr30K and MS COCO (Young et al., 2014; Lin et al., 2014). Both of these data sets contain images with multiple crowd-sourced descriptions per image. These datasets are typically used to train data-driven natural language generation systems to automatically learn to associate visual features with natural language descriptions (Bernardi et al., 2016). Following the training phase, image description systems are evaluated by comparing their output with human generated descriptions for the same image (using textual similarity metrics like BLEU or METEOR, Papineni et al. 2002; Denkowski and Lavie 2014). The standard for what the image descriptions should look like is implicit in the corpus. The only point at which any explicit guidelines are provided is during the crowd-sourcing task, where annotators are given general instructions about what their description should look like. Here are the Flickr30K instructions (the MS COCO instructions are similar):

> 1. Describe the image in one complete but simple sentence. 2. Provide an explicit description of prominent entities. 3. Do not make unfounded assumptions about what is occurring. 4. Only talk about entities that appear in the image. 5. Provide an accurate description of the activities, people, animals and objects you see depicted in the image. 6. Each description must be a single sentence under 100 characters.
> (Hodosh et al., 2013, edited for brevity)

These guidelines leave much of the task open for interpretation by the annotator. For example, it is unclear how the descriptions will be used, or what the target audience is (as pointed out by van Miltenburg et al. 2017). Thus, the underspecified nature of the task invites *variation* and *creativity*. It is important for us to understand the extent of this variation because image description corpora currently set the standard for what an image description should look like.

Earlier work has looked at stereotyping behavior, reporting bias, and the use of negations in image descriptions (van Miltenburg, 2016; Misra et al., 2016; van Miltenburg et al., 2016, 2017), and recently Van Miltenburg et al. (2018) provided an overview of measures to quantify diversity. This paper looks at the variation in the labels used to refer to other people, and presents a taxonomy (based on the Flickr30K dataset) that shows the range of properties that crowd-workers consider in the description process. This taxonomy ranges from physical attributes, such as hair color, to attributes concerning socio-economic status (e.g. *unemployed*).

After discussing related work (§2), we present our method to select person-labels and to categorize (partial) labels into semantic categories (§3).

Proceedings of The 11th International Natural Language Generation Conference, pages 415–420,
Tilburg, The Netherlands, November 5-8, 2018. ©2018 Association for Computational Linguistics

Following this, Section 4 shows the resulting taxonomy, with examples from the Flickr30K dataset. Section 5 discusses our taxonomy in light of the recently published Face2Text dataset (Gatt et al., 2018), and considers the reliability of perceived attributes. We believe these contributions will be useful for practitioners interested in the generation of person-descriptions. Our code and data is publicly available online.[1]

2 Related work

Natural Language Generation researchers tasked with describing other people have mostly been concerned with generating referring expressions *without* visual context, usually for well-known entities (e.g. Castro Ferreira et al. 2016; Kutlak et al. 2016). The closest related work comes from Gatt et al. (2018) and Van Miltenburg (2016).

Gatt et al. (2018) present a dataset of images of human faces, with multiple elicited descriptions per image. They annotated a part of their dataset to estimate how many of the descriptions refer to physical (85%), emotional (44%), or inferred (46%) properties of the subjects depicted in the images. In contrast, the present paper presents a more precise taxonomy, and discusses Gatt et al.'s Face2Text dataset in light of this taxonomy.

Van Miltenburg (2016) used the Flickr30K-entities dataset (Plummer et al., 2015) to cluster entity mentions based on their co-reference to the same entities. We refer to these mentions as *labels*. Their clustering approach yields hundreds of groups of labels referring to similar entities. Here is one of those clusters, relating to FACIAL HAIR:

> beard, goatee, beard and mustache, gray beard, black beard, white beard, red beard, braided beard, gray braided beard, long, white beard, long brown beard, flaming red beard, big beard, short beard, bubble beard, large white beard, thick beard, neatly trimmed beard, scruffy beard, red facial hair

Van Miltenburg (2016) uses this example as anecdotal evidence for the richness of image description data, without further analysis. Our paper aims to provide a deeper analysis of the labels used to refer to other people, by manually categorizing the labels into semantically coherent sub-groups. For example, if we look more closely at the FACIAL HAIR cluster, we can see that these terms include references to the KIND OF HAIR (*beard, goatee, mustache*), the COLOR (*gray, black, white*), LENGTH (*long, short*), SIZE (*big, large*), ORDERLINESS (*neatly trimmed, scruffy*), and PRESENTATION (*braided*). This means that, when asked to talk about an image, people consider at least six different variables just to describe facial hair.

3 Method

We created a taxonomy of the labels used to refer to other people by manually sorting the entity labels into different semantic categories, instead of clustering the labels (as in Van Miltenburg 2016). The advantage of manually sorting the labels is that we have full control over the categories. This makes it possible to make more fine-grained distinctions, and to show the breadth of the label distribution. In this paper, we use the English Flickr30K corpus, focusing on the different ways that crowd-workers describe other people.

3.1 Initial selection

The starting point for our categorization is a list of labels. We compiled this list using the Flickr30K-entities annotations provided by Plummer et al. (2015), and listed all labels that were classed as PEOPLE. After normalization, we found 19,634 unique labels, which is too much to categorize by hand.[2] (It is not possible to crowd-source our categorization task, because the categories are not known beforehand.) Hence we focus our efforts only on the 5,526 labels that end with any of the nouns *girl, boy, woman, man, female, male*, or any of their plural forms.[3] This makes the task more manageable, but it also potentially reduces the variation in the data because the selected labels are more homogeneous. Nevertheless, as we will see in Section 4, we still found a broad range of variation in the labels.

During the categorization task, we found several typing errors, and words unrelated to people-labeling. We addressed these issues by semi-automatically correcting the typing errors, and creating a list of stopwords that were automatically removed from the labels. This further reduced the number of unique labels-to-be-categorized from 5526 to 3401.

[1] https://github.com/evanmiltenburg/LabelingPeople

[2] We normalized the labels by lowercasing them, and removing the characters `@+,&()`.

[3] We applied the same approach to the attributes in the Visual Genome dataset (Krishna et al., 2017), but for reasons of space we focus on Flickr30K. Results are available online.

Category	Examples
Ability	wheelchair bound, able-bodied, disabled, handicapped, blind, one-armed, legless, crippled
Activity	running, chasing, waving, speaking, parachuting, roller-skating, protesting, partying, hiking
Age	young, old, middle-aged, adult, elderly, infant, twenty-something, teen-aged, adolescent
Attractiveness	attractive, beautiful, pretty, sexy, cute, ugly, adorable, hot, handsome, nice, good looking
Build	petite, muscular, slender, lanky, heavy chested, potbellied, well built, burly, stocky, potbellied
Cleanliness	dirty, shaggy, scruffy, muddy, disheveled, messy, well-groomed, grouchy looking, dirty faced
Clothing – amount	shirtless, topless, barefooted, scantily clad, nude, unclothed, undressed, semi-naked, shoe-less
– color	green black uniformed, brightly dressed, red shirted, colorfully clothed, vibrantly colored
– kind	uniformed, casually dressed, sari-garbed, leather-clad, robed, suited, kilted, gothic-dressed
Ethnicity	african-american, asian, oriental, caucasian, chinese, foreign, middle-eastern, indian, tribal
Eyes	blue-eyed, brown eyed, green eyed, bespectacled, glasses-wearing, sun-glassed
Fitness	physically fit, healthy fit, in shape, healthy and fit, weak looking, out-of-shape
Group	cast, circle, audience, crowd, ensemble, couple, team, roomful, group, trio, bunch, gathering
Hair – Color	blond, dark-haired, brown-haired, brunette, redheaded, fair, dark, ginger, dirty-blonde, graying
– Facial	bearded, goateed, shaved, white-bearded, mustachioed, stubbled, green bearded, clean-shaven
– Length	bald, short-haired, long-haired, balding, nearly bald, partially bald, shaved head, bald-headed
– Style	curly-haired, frizzy-haired, pony-tailed, shaggy-haired, curly, dreadlocked, spiky haired
Height	tall, short, petite, taller, long, littler, tall looking, shorter, rather tall, slightly taller
Judgment	stylish, tacky looking, strange, silly, odd looking, hip, comical, flamboyant, shady, shadowy
Mood	happy, excited, curious, enthusiastic, tired, thoughtful, pensive, angry-looking, weary, sad
Occupation	military, navy, photographer, coast guard, executive, cooking professional, bartender
Religion	muslim, hindu, amish, christian, islamic, religious, jewish, buddhist, catholic, mormon, hindi
Social group	homeless, goth, hippie, rasta, peasant, unemployed, poor looking, trash, middle class, high class
State	drunk, extremely drunk, wet, bloody, pregnant, sweaty, cold, handcuffed, ill, injured, deceased
Weight	overweight, fat, slim, skinny, obese, plump, heavyset, heftier, mildly overweight, heavy, hefty

Table 1: Taxonomy of labels referring to other people, with selected examples for each category. All examples are (partial) labels from the Flickr30K dataset.

3.2 Sorting procedure

We manually sorted (partial) labels into semantic categories, shown in Table 1. Nothing crucially hinges on these specific categories, but from our experience with image description datasets, we believe they provide a good first approximation, capturing the breadth of the labels used by the crowd. Our sorting procedure works as follows.

1. Start with a set of labels to be categorized.
2. Remove task-specific stopwords and unrelated phrases (e.g. *a picture of*) from the labels. This reduces the number of unique labels.
3. Select (partial) labels from the list, add them to an existing category file, or create a new category file with those labels.
4. Match the labels with the categories. We use a context-free grammar (CFG, see Figure 1; implemented using the NLTK, Bird et al. 2009) because each label may consist of multiple modifiers from different categories. For example: *African-American young man* has both Ethnicity and Age modifiers.
5. Remove matches from the set of labels to be categorized.
6. Either stop categorization, or go to 3.

Label → Mod, GenderedNoun
Label → Mod, Mod, GenderedNoun
Mod → Ability | Activity | Age | ...
GenderedNoun → woman | man | girl | boy | ...
Age → young | old | middle-aged | adult | elderly | ...
Ethnicity → African-American | Asian | oriental | ...

Figure 1: Subset of our Context-Free Grammar, designed to match labels with different categories of modifiers. Production rules are based on our category files (which are updated in step 3).

Our goal is to get an overview of the different kinds of labels used by the crowd-workers, not to achieve a perfect categorization of all labels. Thus, our stopping criterion is as follows. The sorting task is finished whenever there are no more examples matching existing categories, or warranting new categories. New categories are warranted if there are multiple (partial) labels that clearly fall under the same umbrella, but do not fit into any of the existing categories.

4 Results

We sorted the (partial) labels into 20 different categories, until we were left with only 341 labels

(10%) that could not be fully matched with our categories by the CFG matcher. Examples of uncategorized labels are *birthday girl* and *blood pressure of a man*. The former could be classed as a role associated with an event, but we did not find many such examples. The latter is an artifact of the automated label categorization process for the Flickr30K Entities dataset.

Table 1 shows the 20 different label categories, with examples for each category. With this table, we have an empirically derived taxonomy that provides an overview of the choices that crowd-workers make in order to describe other people. The different categories show the diversity and breadth of the label distribution. In future work, we hope to extend the coverage of our taxonomy (ideally to all 19,634 person-labels in Flickr30K-Entities), and present statistics about the proportion of person-labels from the Flickr30K dataset that fall into each category.

Our taxonomy also provides a reference point to think about the characteristics that we would and would *not* like image description systems to describe. For example, the automatic description of features like RELIGION, WEIGHT, or SOCIAL GROUP would probably do more harm than good. Table 1 also shows us what makes image description difficult. For this domain alone, to produce human-like descriptions, systems need to be able to predict 20 different kinds of features, and decide which feature values are relevant to mention. A further complication is that even after deciding which characteristics to describe, there are still within-category choices to be made. For example, when describing a game of basketball, one might choose to talk about a *man playing basketball* (seeing basketball-playing as a transient property), or *male basketball player* (seeing basketball-playing as an inherent property). These choices go beyond the scope of this paper, but see Beukeboom 2014; Fokkens et al. 2018 for a discussion.

5 Discussion and Future Research

5.1 Extending the taxonomy to Face2text

We obtained the Face2Text corpus (Gatt et al., 2018, v0.1) from the authors to see to what extent our taxonomy could be applied to their data. The main difference between the Flickr30K-Entities labels and the Face2Text descriptions is that the former are part of a larger description, whereas the latter are full-blown descriptions themselves. As a result, the Face2Text descriptions are much longer (a mean of 26.9 tokens versus 2.4 for the Flickr30k-entities labels). This leads to crowd-workers providing much more (and seemingly more specific) information about the people in the images. For example, there are 24 occurrences of 'jaw' in Face2Text, with modifiers such as *angular, pointy, traditional square* to denote the specific shape of the jaw. Such details do not seem relevant enough to mention in a short label, as in the Flickr30K-Entities dataset.

In future work, we hope to extend our taxonomy to cover the Face2Text data. This would make users more aware of the contents of the corpus, and enable them to make a conscious choice about the kinds of features they would like their face description systems to generate.

5.2 Consistency is no substitute for truth

In earlier research, Song et al. (2017) present a system that is able to predict (to varying degrees of success) perceived social attributes from faces. Human participants rated faces from a large database for their attractiveness, friendliness, familiarity, but also to what extent they thought the subjects were egotistical, emotionally stable, or responsible.[4]

It is important to stress that these ratings only indicate *perceived* characteristics, and do not necessarily reflect the actual characters of the individuals in the dataset. More generally, even though people may be able to consistently ascribe a particular property to an individual, this alone does not entail that the property actually applies (see Todorov et al. 2013; Agüera y Arcas et al. 2017 for a discussion). When considering different ways to label other people, we should ask ourselves: is it reasonable to predict this label category based on visual information alone?

5.3 Limitations

The approach taken in this paper has three main limitations, which we will discuss in turn.

First, our taxonomy is based on a subset of the person-labels in the Flickr30K-Entities dataset, and thus may overlook other relevant label categories.

[4] Song et al. (2017) list the following 20 pairs of social traits: (attractive, unattractive), (happy, unhappy), (friendly, unfriendly), (sociable, introverted), (kind, mean), (caring, cold), (calm, aggressive), (trustworthy, untrustworthy), (responsible, irresponsible), (confident, uncertain), (humble, egotistical),(emotionally stable, emotionally unstable), (normal, weird), (intelligent, unintelligent), (interesting, boring), (emotional, unemotional), (memorable, forgettable), (typical, atypical), (familiar, unfamiliar) and (common, uncommon).

We emphasize that our work is not meant to provide an exhaustive categorization of the labels used in the Flickr30K data. Rather, our goal is to highlight the breadth of the label distribution. The fact that the broad taxonomy developed in this paper is based on a subset of all the labels (less than a third of the Flickr30K data) only supports the main point of this paper, which is that humans use a wide array of terms to refer to other people.

Second, our taxonomy is constructed manually, and it is unclear whether replication would yield similar results. This is a natural result of a manual categorization of the person labels, and it would be interesting to see if we could automatically induce a similar taxonomy from the corpus data (for example using LDA; Blei et al. 2003). To facilitate future research in this area, we made all our code and data available online.[1]

Finally, our taxonomy is exclusively based on English, without any input from other languages. It may be the case that speakers of other languages highlight other features, in making reference to other people. This idea opens up another avenue of research, asking two related questions:

1. Do speakers of the same language tend to mention the same person-attributes for the same images?
2. Are there any cross-linguistic differences in what features are mentioned in reference to other people?

Although some work has *mentioned* cross-linguistic differences in how annotators refer to other people (e.g. Li et al. 2016; van Miltenburg et al. 2017), we are not aware of any systematic study that specifically looks at how speakers of different languages make reference to other people, and what features they tend to mention.

6 Conclusion

We have looked at the variation in the ways crowd-workers talk about other people in the Flickr30K dataset. Our main result is that this variation covers a wide range of variables, from appearance to socio-economic status. We formalized this variation in a taxonomy of person-labels, which should help us reflect on the image description task, and the kinds of descriptions that image description systems should produce. Future research should be aware that, even though crowd-workers may systematically produce particular labels, this does not mean that the label is true. We encourage the development of standards and guidelines, that tell us which kinds of labels to use in what kind of situations. Such guidelines may benefit system evaluation and help us avoid the inappropriate labeling of other people.

Acknowledgments

This paper was written while the first author was at the Vrije Universiteit Amsterdam, supported by the NWO Spinoza Prize awarded to Piek Vossen.

References

Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's new clothes. Medium.

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. 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.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Gagestein Sarah, and Wouter Van Attveldt. 2018. Studying Muslim Stereotyping through Microportrait Extraction. In *Proceedings of the Eleventh International Conference on Language Resources and*

Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Albert Gatt, Marc Tanti, Adrian Muscat, Patrizia Paggio, Reuben A Farrugia, Claudia Borg, Kenneth P Camilleri, Mike Rosner, and Lonneke van der Plas. 2018. Face2text: Collecting an annotated image description corpus for the generation of rich face descriptions. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference (LREC'18)*.

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.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73.

Roman Kutlak, Kees van Deemter, and Chris Mellish. 2016. Production of referring expressions for an unknown audience: A computational model of communal common ground. *Frontiers in psychology*, 7:1275.

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.

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

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2017. Cross-linguistic differences and similarities in image descriptions. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 21–30, Santiago de Compostela, Spain. Association for Computational Linguistics.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741. Association for Computational Linguistics.

Emiel van Miltenburg, Roser Morante, and Desmond Elliott. 2016. Pragmatic factors in image description: The case of negations. In *Proceedings of the 5th Workshop on Vision and Language*, pages 54–59, Berlin, Germany. Association for Computational Linguistics.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.

Amanda Song, Linjie Li, Chad Atalla, and Garrison Cottrell. 2017. Learning to see people like people: Predicting the social perception of faces. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Alexander Todorov, Peter Mende-Siedlecki, and Ron Dotsch. 2013. Social judgments from faces. *Current opinion in neurobiology*, 23(3):373–380.

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.