How do image description systems describe people? A targeted assessment of system competence in the PEOPLE domain

Emiel van Miltenburg
Tilburg center for Cognition and Communication
Tilburg University
Warandelaan 2, 5037 AB Tilburg, The Netherlands
C.W.J.vanMiltenburg@tilburguniversity.edu

Abstract

Evaluations of image description systems are typically domain-general: generated descriptions for the held-out test images are either compared to a set of reference descriptions (using automated metrics), or rated by human judges on one or more Likert scales (for fluency, overall quality, and other quality criteria). While useful, these evaluations do not tell us anything about the kinds of image descriptions that systems are able to produce. Or, phrased differently, these evaluations do not tell us anything about the cognitive capabilities of image description systems. This paper proposes a different kind of assessment, that is able to quantify the extent to which these systems are able to describe humans. This assessment is based on a manual characterization (a context-free grammar) of English entity labels in the PEOPLE domain, to determine the range of possible outputs. We examined 9 systems to see what kinds of labels they actually use. We found that these systems only use a small subset of at most 13 different kinds of modifiers (e.g. *tall* and *short* modify HEIGHT, *sad* and *happy* modify MOOD), but 27 kinds of modifiers are never used. Future research could study these semantic dimensions in more detail.

1 Introduction

Image description systems are typically trained and evaluated using datasets of described images, such as Flickr30K and MS COCO (Young et al., 2014; Lin et al., 2014). Automated metrics such as BLEU, Meteor, and CIDEr compare the generated descriptions to a set of reference descriptions, and produce a textual similarity score (Papineni et al., 2002; Denkowski and Lavie, 2014; Vedantam et al., 2015). The overall score is said to convey the performance of the system. These metrics (and especially BLEU) are often criticized because of their low correlation to human ratings (Elliott and Keller, 2014; Kilickaya et al., 2017; Reiter, 2018). Human ratings, in turn, are typically collected using Likert scales, where participants rate descriptions for their fluency, correctness, overall quality, and other quality criteria (van der Lee et al., 2019). While these measures do afford us with some insight into system performance, they are very general, and not tied to any specific cognitive ability (e.g. to produce descriptions containing negations, to reason about past and future events, or to successfully refer to properties or entities in a particular domain). This is a problem because it means we do not know what systems are good or bad at.[1] Perhaps the most informative metric in common use is SPICE (Anderson et al., 2016), which enables researchers to assess the ability of image description systems to produce color terms, for example, by computing precision/recall of these terms with regard to a scene graph representing the relevant image.

Instead of general evaluation metrics, we propose to develop specific metrics to assess[2] whether image description systems are able to produce specific kinds of descriptions. This paper presents a proof-of-concept to assess the extent to which systems are able to produce different kinds of person-labels. Our main idea is that, to assess model performance, we should first characterize the domain of interest, such that we have a clear sense of the range of possible descriptions. Following this characterisation, we can

[1] For a more in-depth discussion of this point, see (Schlangen, 2019; Schlangen, 2020).

[2] Throughout this paper, we deliberately use the term *assessment* instead of *evaluation*, to emphasise that we are interested in characterising model behaviour, rather than determining whether this behaviour is good or bad.

Proceedings of the Second Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 30–36
Barcelona, Spain (Online), December 13, 2020

see whether the model outputs cover the full range (which is unlikely), or whether it produces a small subset of the possible descriptions. Our starting point is the taxonomy of person-labels developed by van Miltenburg et al. (2018b), who manually categorised all person-labels in the Flickr30K (Young et al., 2014) and Visual Genome (Krishna et al., 2017) datasets (both US English), with the head nouns *male, female, males, females, man, woman, men, women, boy, girl, boys* and *girls*. They implemented their categorisation scheme in a context-free grammar (CFG; using NLTK (Loper and Bird, 2002)), so that all combinations of modifiers and head nouns would be covered. To illustrate: with only two modifiers (*tall* and *happy*) and 12 head nouns, there are already 48 possible combinations (24 with single modifiers, and another 24 with both modifiers in two different orders). All variations can be captured by the following CFG rules (where commas are used to indicate a choice between different terminals):

LABEL → DET MOD NOUN	MOD → HEIGHT	MOOD → *happy, sad, angry*
MOD → MOD MOD	DET → *the, a*	HEIGHT → *tall, short*
MOD → MOOD	NOUN → *man, woman, ...*	

The original CFG is limited to the eight head nouns specified above. We extended the grammar to obtain broader coverage of system output data (both in terms of head nouns and in terms of modifiers). We then used the grammar to analyse this data, and to see what kinds of labels are typically produced by image description systems. We found that systems are limited in their coverage: at most 13 out of 40 modifier categories were used. This means that 26 different kinds of semantic properties that are attested in human image description data, cannot be found in the output data.[3]

2 Method

Corpus data. van Miltenburg et al. (2018b) developed two taxonomies using English corpus data from Flickr30K and the Visual Genome dataset (the latter dataset contains 108,077 images from MS COCO). For this paper, we merged the two taxonomies to have broader coverage. After combining the different sets of words for each category, we manually inspected and revised the category files.[4]

System data. We use output data from nine different image description systems, collected by van Miltenburg et al. (2018a), who obtained this data by contacting the authors of all image description papers that appeared in conferences and journals in 2016-2017. All systems were trained o the MS COCO training set, and all descriptions were generated for the MS COCO validation set (40,504 images). Although the systems are slightly older, many are still competitive with the state-of-the-art. Furthermore, we do not want to make any claims about current model performance. Rather, our aim is to provide a proof-of-concept of a new assessment procedure for image description systems.

Preparation. We updated the grammar in two stages. In the first stage, we added more head nouns. People are not just referred to using gendered nouns. Some are referred to using their occupation (*police officer*) or their (social) role (*friend, neighbour, mother*). Thus we added more categories (with corresponding lexical items) for different kinds of head nouns. This required us to determine which nouns refer to people. Because this experiment is a proof-of-concept, we focus only on the nouns that actually occur in the output data for the systems mentioned above. We used SpaCy (Honnibal and Montani, 2017) to identify noun chunks in the generated descriptions, and identified 5385 unique sequences of one or more nouns (assuming that multiple nouns together form compound nouns). We then manually selected 170 heads that refer to people and developed category labels for all different kinds of nouns, adding new labels until all nouns were categorised. A single head may be part of multiple different categories, and thus receive multiple labels. Some examples are provided in Table 1.

In the second stage, we updated the set of modifiers. Since different head nouns may be modified in different ways, we extended coverage of the modifiers to include the ones that are used for nouns other than the twelve original nouns. This was done by identifying all noun chunks that refer to people (i.e. end with one of the heads identified in the previous step), analysing them with our updated grammar, and manually categorising all modifiers that weren't already covered by our grammar. We found that the

[3] All code and data is available at: `https://github.com/evanmiltenburg/AnalysePeopleDescriptions`.

[4] Details about this procedure (with data at each step of the process) are in the GitHub repository.

ACTIVITY	baseball player, commuter, kite surfer, pedestrian, runner, skier, swimmer, spectators
AGE	adult, baby, babies, child, children, kid, toddler, girl, lady, woman, women, man, men
GENDERED	boy, girl, man, females, woman, ladies, bride, groom, daughter, cowboy, camera man
PLURAL OR MASS NOUNS	crowd, passengers, persons, business people, construction workers, students, spectators
RELATION	father, friend, patient, bride, daughter, couple, family, friends, owner, customer, spectators
STATUS OR OCCUPATION	clown, coach, police officer, farmer, king, camera man, students, emergency personnel

Table 1: Different noun categories used in our grammar, with examples for each category.

System	#Labels	#Mods	Avg-mods	Mod-cats	Noun-cats
Tavakoli et al. (2017)	14039	1676	0.12	4	6
Zhou et al. (2017)	14698	1792	0.12	4	6
Vinyals et al. (2017)	15496	1165	0.08	5	6
Shetty et al. (2016)	15247	1514	0.1	4	6
Dai et al. (2017)	14793	1763	0.12	13	6
Mun et al. (2017)	15645	1906	0.12	3	6
Shetty et al. (2017)	14558	1861	0.13	10	6
Liu et al. (2017)	13856	1296	0.09	3	6
Wu et al. (2017)	14486	2060	0.14	5	6

Table 2: Overall statistics for the nine different systems. Columns show the total number of labels referring to humans, the total number of modifiers used, the average number of modifiers per label, the number of modifier categories, and the number of noun categories.

system outputs contain 667 unique noun chunks that refer to people. Of these, 562 (84%) were already covered by the grammar with the new heads included. This meant we had to check 105 noun chunks (16%) for additional modifiers to include in the grammar. We updated the grammar until there were 5 ungrammatical phrases (0.75%) left that were impossible to categorise.[5] During this process, we added three more nominal heads that were initially overlooked ('traffic girl', 'toy boy' and 'mounted police officers'). The largest gains in coverage were not achieved by adding more modifiers, but by adding the UNK token (occurring in 11 chunks), determiners (*their, these, this*, 16 cases), and quantifiers (*some, many, several*, 38 cases).

Analysis. We analysed the data from all nine systems, using our context-free grammar, to see which modifier and head noun categories are most frequent. Figure 1 shows an example analysis, resulting in the modifier categories HEIGHT and MOOD, and the noun category AGENOUN. We assume that each modifier category is used only once per noun chunk. Some labels are ambiguous, resulting in multiple possible analyses. In those cases, we take the union of all possible labels. This does overestimate model performance, so any deeper analysis of system output should inspect the results to find example usages for each category.

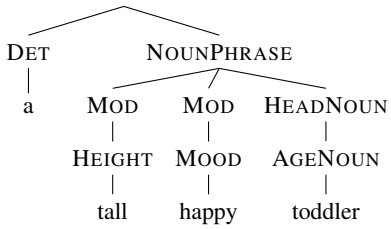

Figure 1: Parse tree for the label: *a tall happy toddler*.

3 Results

Table 2 shows the overall statistics for our corpus of system outputs. We find that the systems differ in the amount of person-labels that they generate. The system from Mun et al. generates 1,789 more person-labels than the system from Liu et al. The system from Vinyals et al. produces the lowest average number of modifiers per person-label. Overall, the systems only use a subset of potential modifier categories: only four to thirteen different kinds of modifiers, while there are 40 different categories in our grammar.

What categories are produced? When we look at the nouns and modifiers that are produced by all systems, we find that all systems produce the same noun types. The story is different for the modifiers, where we find the totals given in Table 3. So, for example, only Dai et al.'s system produced one or more descriptions referring to someone's height: *a short young boy* (analysed as: HEIGHT, AGE,

[5]These were: *a near a crowd, a womans lady, cars police, lawn women, two cut men*

Category	Examples	Total	Systems
AGE	a *baby* boy, a *middle aged* surfer, a pretty *young* man, an *old* lady	9	All
BUILD	a *tiny* baby, a *small* adult/baby/boy/child/girl/kid, a *big* cute little girl	9	All
OCCUPATIONSOCGROUP	the *professional* player/skier, a *little league baseball* player	9	All
ACTIVITY	a *large commercial airplane flying* commuter, a *newly married* couple	6	1,3,4,5,7,9
MOOD	a *happy* boy/child, a *smiling* boy/child/lady/man/woman/young lady	3	5,7,9
ATTRACTIVENESS	a *beautiful* woman/young lady/blond woman, a *cute* baby	3	3,5,7
HAIRCOLOR	a beautiful *blond* woman, a beautiful *blonde* lady, a *blonde* woman	2	5,7
SKINCOLOR	a *dark* woman, a *white* baby, a *white* tennis player, a *yellow* lifeguard	2	5,7
AMOUNTOFCLOTHING	a *naked* baby/child/girl/lady/man/woman, a *shirtless* child/man	2	2,7
NUMBER	a *couple* kids/women, a *few* employees/kids, a *lone* skier/surfer	2	5,7
HEIGHT	a *short* young boy, a *tall* man, the *short* men	1	5
ETHNICITY	an *asian* boy/woman	1	5
JUDGMENT	an *old fashioned man*	1	5
KINDOFCLOTHING	a *hat* adult	1	5

Systems legend: 1: Tavakoli et al. (2017), 2: Zhou et al. (2017), 3: Vinyals et al. (2017), 4: Shetty et al. (2016)
5: Dai et al. (2017), 6: Mun et al. (2017), 7: Shetty et al. (2017), 8: Liu et al. (2017), 9: Wu et al. (2017)

Table 3: Different modifier categories, with examples from the machine-generated output. Modifiers are marked in *italics*. Similar phrases using the same modifier with different nouns are indicated using forward slashes. The last two columns refer to the number of systems producing descriptions containing the relevant types of modifiers.

GENDEREDNOUN/AGENOUN). This might be due to their GAN-based approach, which aims to make descriptions more human-like.

Some observations. Below are some initial observations about the results:

Lack of diversity. The descriptions are not very diverse (van Miltenburg et al., 2018a), and this is reflected in our data: there is little variation within each of the categories. For example, in the BUILD category, only three adjectives (out of 48 possible modifiers) are used: *big, small* and *tiny*.

Overgeneralisation. Our parser overgeneralises. For example, when we take a closer look at the labels containing modifiers relating to the BUILD of a person (e.g., *big, small, tiny*), we find multiple instances where these modifiers are not used in the relevant sense, namely: *a big crowd, a small family, a small team*. This shows that manual analysis is currently still necessary (and the system would clearly benefit from word sense disambiguation), but our parser serves as a good filter: without it, it would not be possible to find different kinds of modifiers at all. With the parser and some manual effort, we have a better picture of the kinds of person-labels a system is able to produce.

Biases. The generated labels seem to be biased with respect to gender and ethnicity (cf. van Miltenburg (2016)). For example: ATTRACTIVENESS modifiers are used almost exclusively with women and children. The only counter-example is ambiguous: 'a *pretty* young man' (pretty young or pretty man). HAIRCOLOR is only specified for women. ETHNICITY is only marked for Asian people.

4 Discussion

(Un)intended use of this work. We intended this work to be used in two ways:

1. As a means to assess the extent to which image description systems are able to produce person-descriptions. This is useful because it allows further investigation of the output: we can use our results to see whether these descriptions are warranted by the images (cf. van Miltenburg (2016)), or investigate whether different kinds of person-labels are fairly distributed across different social groups (cf. Otterbacher et al. (2019)).
2. As a more general example of a targeted assessment procedure. The main takeaway from this paper is that it is possible to assess the ability of image description systems to produce different kinds of PEOPLE-descriptions (and this approach could probably be generalised to other domains).

The generation of person-descriptions is a sensitive area of research, which should be treated with care (Todorov et al., 2013; Agüera y Arcas et al., 2017). Not all of the possible labels in our CFG should be generated by image description systems, because they may be offensive when misapplied, offensive

in general, or simply impossible to predict on the basis of visual features alone. Our work is expressly *not* intended to encourage the production of ever more detailed person-descriptions, since there are clear dual use issues in this area of research.[6,7]

Competence, performance and accuracy. So far we have treated system competence as 'generating descriptions with particular features.' This is a simplification on two counts:

1. It conflates competence and performance (Chomsky, 1965). Systems that don't generate descriptions with particular features for a given set of images, may still be able to generate such descriptions for other images. (Although one might wonder, with the size of the MS COCO validation set, whether those systems would *ever* generate descriptions with the relevant features.)
2. It ignores the accuracy of the generated person-labels, which is essential for being a competent language-user. A full assessment of system competence should include a human evaluation of the different kinds of labels. For example, judges could rate the accuracy of a stratified sample of machine-generated labels from each of the different categories.

Limitations. This work is a proof-of-concept, and our results should be evaluated as such. We have shown that it is possible to characterise an output domain using a context-free grammar, and that we can use such a grammar to assess what kinds of labels are produced by an image description system. Any statistics should be treated as preliminary.

Annotations. Although we took care to double-check all the labels, the taxonomy has been developed by a single annotator. Future work should investigate the replicability of our taxonomy. (Although there are as many taxonomies as taxonomists, and some amount of disagreement is inevitable, there should be a clear overlap between our data and future replications.)

Coverage. Our taxonomy is driven by existing sets of descriptions. Should image description systems become more creative in the future, or rely on different datasets, then this will affect the coverage of our taxonomy, and updates will be necessary. We do believe that this process could be automated, and that it is feasible to update the taxonomy in a matter of days (which we believe is acceptable, and proportionate to the development time of a new image description system).

Ambiguity. As mentioned in the results section, a context-free grammar is not powerful enough to disambiguate different word senses. So although our parser can act as a filter to select relevant noun phrases from system outputs, this will result in false positives for different modifiers. To reduce this issue, one could either refine our taxonomy and tie it to the output of word sense disambiguation systems, or manually correct the outputs of our parser to create a training set, turning description categorisation into a sequence labeling problem. The latter solution would hopefully also generalise to unseen nouns and modifiers, reducing coverage issues caused by our grounded-in-data approach.

5 Conclusion

Our work shows that it is feasible to assess how image description systems label people, using a context-free grammar. This idea could be applied more broadly, to assess systems' coverage of the original training data (i.e. which patterns the system managed to capture). The development of a taxonomy of expressions in a particular domain does represent a serious time investment, but once this is done, we can more closely assess system performance in a particular domain, for any number of systems. (And perhaps this is also just the cost of wanting to develop end-to-end systems using crowdsourced image description corpora.) As noted by van Miltenburg et al. (2018b), developing a taxonomy also forces us to think about the desirability of different kinds of expressions that may be used in the generated descriptions. This will hopefully lead to improved image description guidelines, which may later be formalised in an assessment procedure, similar to the one reported in this paper.

[6] Dual use refers to the idea that technology can be used both for good and for nefarious purposes (Hovy and Spruit, 2016).

[7] On a related note, visually impaired technology advocate Chancey Fleet has said that she "[worries] that the beneficence that kind of surrounds accessibility technology, this idea we have that accessibility is always a universal good, has created a situation where blind people and our perceived needs can be used to normalize facial recognition in private and semiprivate spaces" (Morris et al., 2020). Whenever researchers believe computer vision technology may help visually impaired users, it is necessary to confirm with these people whether they would indeed appreciate this technology, and whether they believe it is worth any potential negative implications, such as increased surveillance and a further invasion of privacy.

References

Blaise Agüera y Arcas, Margaret Mitchell, and Alexander Todorov. 2017. Physiognomy's new clothes. Medium, May 17.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 382–398, Cham. Springer International Publishing.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the 2017 International Conference on Computer Vision*, pages 2970–2979, Venice, Italy.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland, June. Association for Computational Linguistics.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear. See `https://spacy.io` or `https://github.com/explosion/spaCy`.

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.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain, April. Association for Computational Linguistics.

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. *International Journal of Computer Vision*, 123(1):32–73.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Chang Liu, Fuchun Sun, Changhu Wang, Feng Wang, and Alan Yuille. 2017. Mat: A multimodal attentive translator for image captioning. In *IJCAI*, pages 4033–4039.

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.

Meredith Ringel Morris, Cynthia Bennett, Chancey Fleet, and Venkatesh Potluri. 2020. Vizwiz grand challenge workshop at cvpr 2020 - panel discussion with blind technology experts. Video available through YouTube: `https://www.youtube.com/watch?v=f613diLbVAc`, workshop address: `https://vizwiz.org/workshops/2020-workshop/`.

Jonghwan Mun, Minsu Cho, and Bohyung Han. 2017. Text-guided attention model for image captioning. In *AAAI Conference on Artificial Intelligence*.

Jahna Otterbacher, Pınar Barlas, Styliani Kleanthous, and Kyriakos Kyriakou. 2019. How do we talk about other people? group (un)fairness in natural language image descriptions. In *Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing*, pages 106–114. Association for the Advancement of Artificial Intelligence (AAAI).

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

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401, September.

David Schlangen. 2019. Language tasks and language games: On methodology in current natural language processing research. *CoRR*, abs/1908.10747.

David Schlangen. 2020. Targeting the benchmark: On methodology in current natural language processing research. *CoRR*, abs/2007.04792.

Rakshith Shetty, Hamed R.-Tavakoli, and Jorma Laaksonen. 2016. Exploiting scene context for image captioning. In *Proceedings of the 2016 ACM Workshop on Vision and Language Integration Meets Multimedia Fusion*, iV&L-MM '16, pages 1–8, New York, NY, USA. ACM.

Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*, Oct.

Hamed R Tavakoli, Rakshith Shetty, Ali Borji, and Jorma Laaksonen. 2017. Paying attention to descriptions generated by image captioning models. In *CVPR*, pages 2487–2496.

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

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368, Tokyo, Japan, October–November. Association for Computational Linguistics.

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

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018b. Talking about other people: an endless range of possibilities. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 415–420, Tilburg University, The Netherlands, November. Association for Computational Linguistics.

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.

R. Vedantam, C. L. Zitnick, and D. Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2017. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, April.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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.

Luowei Zhou, Chenliang Xu, Parker Koch, and Jason J Corso. 2017. Watch what you just said: Image captioning with text-conditional attention. In *Proceedings of the ACM International Conference on Multimedia Thematic Workshops*.