Gradations of Error Severity in Automatic Image Descriptions

Emiel van Miltenburg[1] ✉, **Wei-Ting Lu**[1], **Emiel Krahmer**[1], **Albert Gatt**[2], **Guanyi Chen**[3], **Lin Li**[3], and **Kees van Deemter**[3]

[1]Tilburg center for Cognition and Communication (TiCC), Tilburg University
[2]Institute of Linguistics and Language Technology, University of Malta
[3]Department of Information and Computing Sciences, Utrecht University
✉ Corresponding author: C.W.J.vanMiltenburg@uvt.nl

Abstract

Earlier research has shown that evaluation metrics based on textual similarity (e.g., BLEU, CIDEr, Meteor) do not correlate well with human evaluation scores for automatically generated text. We carried out an experiment with Chinese speakers, where we systematically manipulated image descriptions to contain different kinds of errors. Because our manipulated descriptions form minimal pairs with the reference descriptions, we are able to assess the impact of different kinds of errors on the perceived quality of the descriptions. Our results show that different kinds of errors elicit significantly different evaluation scores, even though all erroneous descriptions differ in only one character from the reference descriptions. Evaluation metrics based solely on textual similarity are unable to capture these differences, which (at least partially) explains their poor correlation with human judgments. Our work provides the foundations for future work, where we aim to understand *why* different errors are seen as more or less severe.

1 Introduction

Recent years have seen a growing discomfort with the use of automatic metrics like BLEU (Papineni et al., 2002) for the evaluation of natural language generation (NLG) systems (e.g., Sulem et al. 2018; Reiter 2018; Mathur et al. 2020). Much of the criticism centers around the fact that these metrics show poor agreement with human judgments. While many researchers have tried to develop new metrics that are better suited to evaluate NLG systems (e.g. tailored to the domain like SPICE (Anderson et al., 2016) or with intensive pre-training like BLEURT; Sellam et al. 2020), we are not aware of any studies attempting to *explain* why we see such a poor correlation between human judges and automatic metrics. This paper aims to explore this hypothesis, focusing on the evaluation of automatic image description systems. We focus on image descriptions because it is relatively easy for humans to judge whether a given description correctly describes an image. Compare this to the WebNLG data (Gardent et al., 2017), where participants would have to judge whether a given sentence verbalizes a set of triples containing information about properties of different entities, and how different entities relate to each other. The format of the input data, as well as the numerous ways to verbalise each triple separately, and express them jointly, perhaps through some process of aggregation, is bound to make the judgment more challenging.

1.1 Motivation

Image description systems make different kinds of mistakes, and these mistakes are likely to be of different importance for a 'correct' interpretation of the relevant image. Consider Figure 1, which shows multiple human reference descriptions, and a description generated by Li et al.'s (2018) system (all in Chinese, with English glosses). This system makes three different mistakes, which are shown separately in Example (1; edited for brevity). We refer to these mistakes as an *age error* (1b), *color error* (1c), and an *object error* (1d).[1]

(1) Gold standard (a) and errors (b–d) from Fig. 1.

a. A woman wearing a blue shirt holds a cake.
b. A **girl** wearing a blue shirt holds a cake.
c. A woman wearing a **red** shirt holds a cake.
d. A woman wearing a blue shirt holds a **racket**.

Intuitively, the different errors made by the system are not equally severe. Our intuition is that the object error is much more flagrant than the age

[1]Note that the human annotators do not make any mistakes at all; they are clearly able to identify the protagonist as a woman who is wearing a blue shirt and holding a cake.

Proceedings of The 13th International Conference on Natural Language Generation, pages 398–411,
Dublin, Ireland, 15-18 December, 2020.

Human: 戶外 一個 穿著 藍色 恤 的 女人 手 裡 拿著 一 盤子 五彩 蛋糕
outdoor one wear blue shirt of woman hand in take one plate five-color cake
'A woman wearing a blue shirt outside, has a plate with a five-colored cake in her hand'

一個 女人 端著 兩 塊 彩虹 蛋糕
one woman hold two piece rainbow cake
'A woman holding two pieces of rainbow cake'

一個 女人 拿著 盤子 上面 有 兩 塊 彩色的 蛋糕
one woman take plate on is two piece rainbow cake
'A woman holding a plate with two pieces of rainbow cake'

一個 女人 端著 兩 塊 生日 蛋糕
one woman hold two piece birthday cake
'A woman holding two pieces of birthday cake'

一個 女人 拿著 放 有 彩虹 蛋糕 的 盤子
one woman take put is rainbow cake of plate
'A woman holding a plate of birthday cake'

System: 一個 穿著 **紅色** 上衣 的 **女孩** 手 裡 拿著 一個 **網球拍**
one wear **red** shirt of **girl** hand in hold one **tennis racket**
'A girl wearing a red shirt is holding a tennis racket'

Figure 1: Image 59547 from the MS COCO dataset, with human descriptions from the COCO-CN corpus, and an automatically generated description from Li et al.'s system. Glosses are ours. Errors in the automatically generated description are **in bold and highlighted in red**. Original picture taken by Matt Bauer (CC BY-NC 2.0).

error. A potential explanation for this difference lies in the fact that AGE is a more *vague* property than OBJECT CATEGORY. We will discuss this idea in Section 3, where we posit our hypotheses.

1.2 Main questions

We address three research questions: 1. Do people's quality judgments indeed differ between error categories? 2. If there is a gradation of error severity, how is it ordered? 3. What might explain those differences? As our title suggests, we indeed find differences in perceived error severity between different types of errors:

1. Perceived severity for GENDER ERRORS is significantly worse than AGE errors.
2. CLOTHING COLOR errors are significantly worse than CLOTHING TYPE or AGE errors.

We discuss potential explanations first in our hypotheses section (§3), and later take stock in the discussion (§6). Although our study uses Chinese-speaking participants, we believe our main result (differences in perceived error severity) should generalize to other languages, though the order of the error categories on the 'severity scale' may differ. We will discuss this issue further in Section 6.6.

1.3 Implications

This paper provides evidence that there are differences in perceived error severity between different kinds of errors in image descriptions. Our results offer one reason why many automatic evaluation metrics correlate poorly with human judgments: most metrics wrongly assume that there is no difference between different kinds of mistakes. This means that we have to rethink the relation between accuracy and overall quality (as measured through human judgments). We will discuss this issue further in Section 6.2.

2 Background

2.1 Errors in NLG output

NLG output is not perfect. Van Deemter and Reiter (2018) discuss how errors may arise at different stages of the NLG pipeline. Much has been written about how to best evaluate the quality of automatically generated text (e.g. van der Lee et al. 2019; Celikyilmaz et al. 2020), but less is known about the impact of different kinds of errors on users of automatically generated text. To our knowledge, responses to errors have only been studied by researchers in Human-Computer Interaction (e.g., Abdolrahmani et al. 2017) or Human-Robot Interaction (e.g., Mirnig et al. 2017). Together, these studies show that while some errors may make users abandon a product, other errors may not be judged as harshly. In fact, Mirnig et al. found that people may even like a robot *more* if it occasionally makes a mistake. But, as Abdolrahmani et al. note: this all depends on the context of use.

Our study asks how we can systematically study the impact of different kinds of errors in automatic image descriptions. Several studies have proposed different categorizations of these errors. We will discuss those studies below.

2.2 Weaknesses in system competence

Hodosh and Hockenmaier (2016) and Shekhar et al. (2017) both manipulate existing image descriptions to generate flawed descriptions, which they use to see if automatic image description systems can recognize those flaws. For example, given a sentence like (2), Hodosh and Hockenmaier swap the existing scene description for another one (2→2a), and ask systems to identify the correct description. Shekhar et al. change an entity with another entity falling under the same supercategory (e.g. VEHICLE, 2→2b), and ask systems to identify the flaw in the description.

(2) Ref: A man is riding a bicycle down the street.

a. A man is riding a bicycle **on the beach**.
b. A man is riding a **motorcycle** down the street.

Together, these studies show that image description systems still have difficulties identifying GRAMMATICAL SUBJECTS AND OBJECTS, SCENES, and OBJECTS in general. An interesting property of the flawed descriptions generated by Shekhar et al. (2017) is that their manipulations are associated with different semantic categories (ANIMAL, VEHICLE, FURNITURE, ...). This enables them to pinpoint which kinds of entities are easier or harder for systems to describe.

2.3 Errors in system output

Anderson et al. (2016) propose the SPICE-metric, which differs from other evaluation metrics in that it uses the reference descriptions to build an abstract scene graph. The hypothesis is also parsed into an abstract scene graph, and compared to the reference graph. These graphs can be represented as tuples that correspond to different features, namely: OBJECT, RELATION, and ATTRIBUTE (which is subdivided into COLOR, COUNT and SIZE). We can use SPICE to identify different kinds of propositions that are communicated by an image description. For example, according to SPICE, sentence (3) conveys five propositions: 1. there are eggs (OBJECT); 2. there are three of them (ATTRIBUTE: NUMBER); 3. there is a basket (OBJECT); 4. the basket is green (ATTRIBUTE: COLOR); 5. the eggs are in the basket (RELATION).

(3) Ref: There are three eggs in the green basket.

a. There are **four** eggs in the green basket.
b. There are three eggs **under** the green basket.

SPICE is not able to determine whether any proposition is correct or not (and thus it does not exactly identify errors), but instead it returns an F1-score over the different propositions, showing how often systems 'retrieve' the same propositions that can be extracted from the reference data. For our purposes, we can use the SPICE categories to reason about error severity. For example, intuitively, COUNT errors (such as 3a) might be more forgivable in the eyes of human judges than RELATION errors (such as 3b), since it's easy to miscount.

To our knowledge, Van Miltenburg and Elliott (2017) provide the most extensive error analysis of automatic image descriptions. In addition to the semantic categories identified by Anderson et al. (2016), they discuss: POSITION and ACTIVITY; more attributes: AGE, GENDER and STANCE (e.g. whether someone stands, sits, or crouches); SCENE/EVENT/LOCATION; and different ways to generate the wrong subject or object: confusing it with a similar entity, hallucinating an entity, identifying the wrong entity for the semantic role, or identifying the correct entity but wrongly adding another subject/object. Finally, the authors observe that there are surprisingly many errors concerning TYPE OF CLOTHING and COLOR OF CLOTHING.

3 Hypotheses

In line with the error analysis from van Miltenburg and Elliott (2017), our experiment explores the perceived severity of four common kinds of errors found in automatic image description systems, relating to 1. AGE, 2. GENDER, 3. CLOTHING-COLOR, and 4. CLOTHING-TYPE. This section discusses our expectations regarding the quality scores for sentences containing these types of errors.

Earlier studies in linguistics have shown that not every word in a sentence is equally prominent (see Lockwood and Macaulay 2012; Himmelmann and Primus 2015 for an overview). For example, the subject of a sentence is more prominent than the direct object, which in turn is more prominent than the indirect object. By the same token, certain expressions may achieve prominence due to the type of entity they denote. For example, people may be more prominent than inanimate objects (like clothes); an observation also borne out by studies on how humans process visual inputs (cf.

Yun et al. 2013). Animacy also plays an important role in referring expression generation (Baltaretu et al., 2016; Vogels et al., 2013). We believe that animacy might also play a (yet to be determined) role in quality judgments, and our intuition is:[2]

Hypothesis 1: The perceived quality of descriptions with people-related errors is lower than the perceived quality of descriptions with clothing-related errors.

Next our hunch is that, in most situations, gender errors are worse than age errors, and errors regarding clothing type are worse than errors regarding clothing color. Two perspectives come to mind:

1. Function. Clothing type is a more essential property of a piece of clothing than its color. For example, the most important aspect of a T-shirt is that you can wear it to keep your chest covered and warm. Color is secondary.

2. Degrees of vagueness. Expressions in natural language are often *vague*, meaning that they allow for situations in which it is debatable whether the expression has been used truthfully or not (e.g., Williamson 2002; Van Deemter 2012). Color terms are famously vague (e.g., Parikh 1994), because while there may be situations where everyone agrees that X is *red*, the exact boundaries of REDNESS cannot be given.

While it can be argued that all categories exhibit some degree of vagueness, AGE-denoting expressions tend to be more vague than GENDER-denoting ones (e.g. whether someone is old is more often debatable than whether someone is male); likewise, COLOR-denoting expressions tend to be more vague than CLOTHING-TYPE-denoting ones. We therefore expect that participants are more forgiving when judging the truthfulness of age-denoting and clothing-color denoting expressions than when judging the truthfulness of gender-denoting or clothing-type denoting ones. This difference in severity may also become entrenched, so that one type of error may generally be perceived as worse than another. Our intuitions are as follows:

Hypothesis 2: The perceived quality of descriptions with an age error is higher than the perceived quality of descriptions with a gender error.

Hypothesis 3: The perceived quality of description with clothing color error is higher than the perceived quality of descriptions with clothing type error.

[2]Sentence structure is a potential confound in our design. People-related terms (*woman,boy*) and clothing-related terms (*pink, coat*) are both in subject position, but the latter are generally more deeply embedded in the NP [woman [wearing [a [pink coat]]]], making them syntactically less prominent (and so errors may be less obvious).

4 Method

We provide a detailed description of our method below. All data and stimuli are provided in the supplementary materials.

4.1 Participants

We used network sampling to recruit 61 volunteers (35 female, 26 male; 59 native, 2 fluent speakers of Chinese) to participate in our study.[3] Most (N=38) received a university education. All participants indicated that they were not color-blind.

4.2 Materials

We selected 7 images from MS COCO. For each image, we manually constructed four descriptions with exactly one error in each of them, resulting in 28 image-description pairs. Figure 2 shows an example image with the reference description, and four erroneous descriptions.

Image selection. Images from the MS COCO dataset were selected to fit the following criteria:

1. They should be full-color images.
2. There should be a human protagonist, with their face and at least half their body visible.
3. The content of the images should be clearly recognizable.
4. Each clothing item should have a single color.
5. Clothing items should have different colors.

We established these criteria to avoid error ambiguity. For example, if the man in Figure 2 were wearing yellow shorts as well, then the clothing type error could be resolved in two ways: *coat→shirt* or *coat→shorts*. This is undesirable, since differences in error ambiguity may introduce additional variance in our experiment.

Descriptions. The descriptions were written by a native speaker of Chinese, who was tasked to create minimal pairs between the erroneous descriptions and a single reference description. We used four different types of errors: GENDER, AGE, CLOTHING TYPE, CLOTHING COLOR. We discuss our motivation for these categories in Section 3.

As Figure 2 shows, the erroneous descriptions only differ in one Mandarin character from the reference description. This is essential, so that automatic evaluation methods give each erroneous description the same score.[4] To illustrate, Exam-

[3]Although we did not ask for their nationality (and thus cannot provide counts), most participants are Taiwanese.

[4]This is a conservative choice, because Chinese words may consist of one or more (often two) syllables/characters. Assuming the descriptions would be tokenised (i.e. segmented

Correct	一位	男人	穿著	黃色	上衣	在	網球場	打	網球
	A	man	wear	yellow	shirt	on	tennis court	play	tennis
Translation	‘A man in a yellow shirt plays tennis on the tennis court.’								
Gender error	一位	**女人**	穿著	黃色	上衣	在	網球場	打	網球
	A	**woman**	wear	yellow	shirt	on	tennis court	play	tennis
Age error	一位	男**孩**	穿著	黃色	上衣	在	網球場	打	網球
	A	**boy**	wear	yellow	shirt	on	tennis court	play	tennis
Clothing type error	一位	男人	穿著	黃色	**大衣**	在	網球場	打	網球
	A	man	wear	yellow	**coat**	on	tennis court	play	tennis
Clothing color error	一位	男人	穿著	**紫色**	上衣	在	網球場	打	網球
	A	man	wear	**purple**	shirt	on	tennis court	play	tennis

Figure 2: Correct reference description, along with systematically manipulated descriptions for image 344149 from the MS COCO dataset. Each erroneous description differs only in one character from the original. The picture (of Jarkko Nieminen at the 2010 French Open) was taken by JanJan de Paris (CC BY-SA 2.0).

ple (4) presents a Chinese noun formation paradigm for gendered person descriptions:[5]

(4) a. 男人 ‘man’ (MALE+PERSON)
b. 女人 ‘woman’ (FEMALE+PERSON)
c. 男孩 ‘boy’ (MALE+CHILD)
d. 女孩 ‘girl’ (FEMALE+CHILD)

As with the images, we aimed to avoid error ambiguity. For example, suppose that the man in Figure 2 were erroneously referred to as *wearing yellow shorts*. We could resolve this issue in two ways: (1) resolve the color: *black shorts*, (2) resolve the clothing: *yellow shirt*. Because it is not clear which error type is applicable, these kinds of ambiguities would make it impossible to determine the impact of individual error types. Therefore, we ensured that there is always a single fix with the lowest edit distance. Finally, there is likely to be some variance within each error type. For example: in the COLOR ERROR category, the mistake *orange→red* is less severe than *orange→blue*. To minimize this issue, and to focus on between-category differences, we aimed to generate clear-cut examples for each error category. We leave within-error variation for future research.

4.3 Design

Our experiment was implemented in Qualtrics, and followed a within-subjects design, where each participant was exposed to all 28 stimuli (i.e., all images, with all erroneous descriptions). In each trial, participants rated the quality of the erroneous description on a continuous scale from 0 (worst) to 100 (best), using a slider (cf. Magnitude Estimation (Stevens, 1975; Bard et al., 1996), or Direct Assessment; Graham et al. 2018). The erroneous description was always presented in the context of the image and the correct reference description.

4.4 Procedure

Participants were invited to take part in the online experiment through different social media channels. After clicking the Qualtrics link, they were first shown an introductory text with a description of the study (including its aim: to understand how users respond to automatically generated text) and a consent form.[6] After consenting to the study, participants were directed to the trial phase, demographic questions, and the main experiment.

Trial phase. The trial phase consisted of four questions for participants, similar to Figure 3, where they were asked to indicate the quality of the automatically generated description on a slider bar. The purpose of these questions is for the participants to calibrate their responses.

Demographic questions. Participants were asked to indicate their age, gender, education level, Chinese proficiency, and whether they are colorblind or not. We excluded two participants based on these questions: one colorblind participant, and one Chinese beginner.

Main experiment. The main experiment featured the same kind of questions as in the trial phase.

at the word level) first by any evaluation measure, it would also have been defensible to change multiple characters.

[5]Chinese crucially differs from English in that 女孩 (‘girl’) cannot be used to refer to adults, whereas English does allow for ‘girl’ to refer to an adult woman, in colloquial use. Other languages, like Maltese, also pattern with Chinese in this regard. We would expect this kind of age error to be perceived as less severe in English (if it is perceived as an error at all).

[6]Participants agreed to take part in the experiment, and to have their (anonymized) responses recorded and shared with the scientific community. They were informed that they could quit at any time, without any negative consequences.

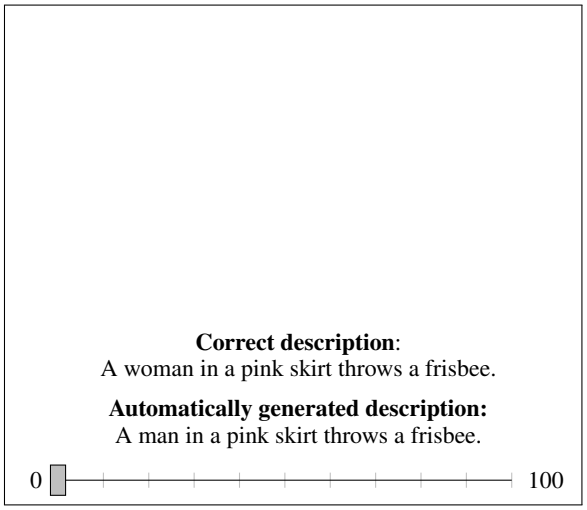

Figure 3: Example item, with the picture (137767 in MS COCO), the reference description, the erroneous description, and a slider to indicate the description quality. Descriptions have been translated and edited for ease of presentation. Original picture taken by Mike LaCon (CC BY-SA 2.0).

Category	Mean	Standard deviation
Age	50.6	23.1
Gender	41.0	23.4
Clothing color	36.5	24.6
Clothing type	45.9	21.5

Table 1: Descriptive statistics for each of the error categories. Mean scores are on a scale from 0–100, where 0 is bad and 100 is good.

Each participant was asked to rate the quality of all 28 stimuli, presented in random order.

Before running our study, we carried out a pretest to get feedback, and to determine the duration of our experiment (5-6 minutes), to inform the participants before taking part in the study.[7]

5 Results

We found that different error types are indeed judged differently; A repeated measures ANOVA revealed a significant overall effect of error type ($F(2.33, 139.5)=13.827$, $p=<0.001$, $\eta^2= 0.05$). Table 1 provides descriptive statistics, showing the different mean scores and their standard deviations.

5.1 Hypothesis evaluation

We subsequently carried out multiple paired sample t-tests to find out which error types significantly differed from each other. The results for these tests are provided by Table 2, and are discussed below.

Hypothesis 1. We expected that people-related errors would be rated worse than clothing-related errors. This is clearly not the case: descriptions containing age errors are significantly better than those with clothing color errors. Errors regarding clothing type seem to be roughly on the same footing as age- and gender-related errors.

Hypothesis 2. We also expected that the perceived quality of descriptions with age errors would be higher than that of descriptions with gender errors. We found that this is indeed the case: scores for age errors are significantly better than the scores for gender errors.

Hypothesis 3. Finally, we expected that clothing type errors would be worse than clothing color errors, but in fact we found the opposite: clothing color errors are significantly worse than clothing type errors.

5.2 Exploratory analysis

Our main analysis revealed significant differences *between* descriptions with different error types. We then looked at differences *within* different error categories. Specifically, we investigated the *direction* of the errors for two error types: (1) Gender: changing male to female (e.g. *man→woman*), versus female to male. (2) AGE: changing young to old (e.g. *boy→man*), versus old to young. Descriptive statistics are provided in Table 3. The means for both gender-related errors are similar, and we failed to find a significant effect of error directionality for gender ($t(60)=0.835$, $p=0.407$).

We did, however, find a significant effect of error directionality for age ($t(60)=-4.49$, $p<0.001$). Changing the label from old (e.g. *man*) to young (e.g. *boy*) on average leads to a 9-point reduction in description quality (on a scale from 0 to 100). This might be due to a difference in error severity, but perhaps a more plausible explanation is that the Chinese classifier 位 is used to express politeness (Huang, 2017). Maybe our participants found it odd to be using this marker with children (e.g., 一位男童) instead of adults (e.g., 一位男人). Unfortunately we cannot know this for sure, because all our stimuli start with the same classifier.

[7] Having a short study and communicating the duration should reduce the dropout rate of our experiment.

Category 1	Category 2	t	df	p-value	Adjusted p-value	Significant?
Age	Clothing color	5.593	60	5.81e−7	3.49e−6	Yes
Age	Clothing type	2.161	60	0.035	0.208	No
Age	Gender	4.739	60	1.36e−5	8.16e−5	Yes
Clothing color	Clothing type	−4.993	60	5.43e−6	3.26e−5	Yes
Clothing color	Gender	−1.680	60	0.098	0.589	No
Clothing type	Gender	2.038	60	0.046	0.276	No

Table 2: Results of multiple paired sample t-tests to compare the means of the scores for the different error categories. The table shows both the original p-values and the Bonferroni-adjusted p-values that were used to determine significance at $\alpha = 0.05$.

Category	Direction	Mean	SD
Gender	Male to female	40.508	23.300
Gender	Female to male	41.601	25.084
Age	Young to old	58.475	23.252
Age	Old to young	49.226	25.748

Table 3: Descriptive statistics for subcategories of AGE and GENDER-related errors. Higher score means greater perceived quality.

6 Discussion

6.1 Explaining our results

Our main finding, that different error types also differ in severity, suggests that people attach different levels of importance to different aspects of an image description. Our hypotheses provided a first attempt at an explanation, although it is clear that more work is needed to develop a better understanding of why these differences in severity arise. For example, we severely underestimated the severity of color errors. In hindsight, we believe that the severity may be related to the fact that color is a prominent feature, and the blatant color errors in our experiment were perceivable at a glance.

Our (revised) intuition is that, similar to visual attention, error severity may be determined both by bottom-up and top-down factors (Itti and Koch, 2000; Borji and Itti, 2013). Some errors (like our color errors) are easily perceived, and may thus elicit strong responses. Others (such as gender errors) may not be as easily perceived, but their social relevance similarly elicits strong responses.

6.2 NLG and risk-taking

Carletta and Mellish (1996) discuss risk-taking in task-oriented dialogue. They show that efficient communication requires speakers to make assumptions about the hearer, and to risk being misunderstood. This is better than the alternative, which is to confirm that all the requisite knowledge (to understand the utterance) is in place. A similar view has been expressed by Clark (1996): interlocutors can rely on various heuristics to make communication more efficient, and obviate the need for exhaustive checking of common ground. Our work can be seen as extending this risk-taking literature, quantifying *the impact of being wrong*.

Every additional detail you provide in a generated text may make the text more useful. But, at the same time, every additional detail you provide carries the risk of being wrong about that detail. Thus there is a trade-off between accuracy and usefulness. Ideally, this trade-off should be resolved by assessing the impact of our decisions. In other words: we should now be able to quantify (1) the usefulness of generating a particular detail, (2) the risk of being wrong about that detail, and (3) the potential impact of being wrong about that detail. We only focused on the latter, showing that different kinds of errors may be rated differently by end users. The risk of being wrong may be approximated through the model's confidence scores. The usefulness of generating a particular detail is (partly) context-dependent.

6.3 Task effects and generalisability

The present paper sought to maintain a 'task-neutral' stance, requiring only that participants rate descriptions in terms of their accuracy with respect to the image. It is however likely that perceived error severity would be strongly impacted by the communicative setting in which a text was being generated. To take an example, our findings suggested that colour errors are more prominent for speakers than initially assumed. Above, we hinted that this could be a largely bottom-up salience effect (Itti and Koch, 2000) due to the contrastive nature of colour in our items.

However, task demands and top-down expectations may make other features more prominent and may also impact the amount of risk-taking a system

or human is willing to take. As an example, consider the setting of the VizWiz challenge (Gurari et al., 2018), which consists of questions asked by visually impaired people seeking help from online users, based on photographs usually taken using phones. Answering a question to help a user find their medication might motivate more detail, less risk, and a stronger reliance on features a system has high confidence in.

While task demands are likely to change the severity of errors, we would still contend that the general point being made is an important one, namely, that not all errors are equally severe. This has implications for our use of evaluation metrics.

6.4 Implications for different metrics

This paper provides a conceptual argument against the use of superficial metrics like BLEU, that only look at textual similarity. In our experiments, erroneous descriptions differ by only one character, so the edit distance is always the same. Differences in perceived severity of different error types thus cannot be explained by these kinds of metrics. A separate weakness of BLEU is that an image may be described by many different descriptions; with a finite amount of references, BLEU may penalise descriptions that provide yet another perspective on the same image. Our results show that BLEU and similar metrics are insufficient even with an infinite amount of different (but correct) references. For SPICE, similar limitations should apply: if the different propositions identified by SPICE are not weighted by the kind of proposition, then this uniform approach will not be able to capture differences in severity. Beyond image description, similar issues may arise in other metrics. For example, referring expression generation systems are often evaluated using DICE (Dice, 1945). Even though this metric looks at meaning (not syntactic form), two referring expressions, RE_1 and RE_2, can obtain the same DICE score yet RE_1 may be intuitively much better than RE_2. The reason here is that DICE looks exclusively at the degree of overlap between the set of properties expressed in a gold standard item and the set of properties expressed by a referring expression produced by a referring expression generation algorithm.

6.5 Vagueness and gradability

Our study relied on a setup where vagueness and gradability have little impact. For instance, the age differences considered are broad enough to make terms such as 'girl' or 'woman' clear-cut (modulo linguistic differences; see below). On the other hand, descriptions may contain gradable terms whose boundaries are debatable (e.g. 'toddler' versus 'baby') and whose usage is harder to classify as an 'error'. The visual input may itself be ambiguous: e.g. it may not be clear whether a person in a photograph is an adult. A plausible hypothesis would be that users would be more tolerant of terms used in borderline cases, or visually ambiguous ones, than more clear-cut cases.

6.6 Future research

Different languages. We only looked at Chinese image descriptions. Our intuition is that other languages show similar gradations in perceived severity of different kinds of errors, but this remains to be tested, paying attention to cross-linguistic differences (such as those noted in Footnote 5 above). This is especially important because our intuitions (even as native speakers) may not be reflected by the data, as we've seen with our results. Another question is whether the average severity of different error categories is similar across different languages. Future research should investigate this question using a typologically diverse sample of languages (cf. Bender 2011).

Different types of errors. We also restricted ourselves to four different types of errors. Future research should look into other kinds of errors, to better understand how different kinds of errors affect the perceived quality of the output. The inclusion of other error categories would also allow us to test hypotheses about the importance of different properties for the representation of visual scenes.

Anscombe's quartet for NLG. Anscombe's quartet is a well-known collection of four datasets that have similar descriptive statistics (mean, variance, correlation of x and y), but that have wildly different plots when you visualize the data (Anscombe, 1973). Our dataset is designed as a linguistic analog to Anscombe's Quartet: all erroneous descriptions differ the same, minimal amount (one character) from the reference description, but we hypothesized them to have very different quality ratings. Analogously to Anscombe's quartet, metrics like BLEU are unable to capture any differences in perceived quality of the descriptions. We encourage NLG researchers to develop similar datasets, so as to put evaluation metrics to the test, to see if they can truly capture differences in perceived quality.

Weighted quality metrics? Given that metrics like BLEU do not correlate well with human judgments (Reiter, 2018; Mathur et al., 2020), and seeing that human judgments are influenced by error types, one might conclude that we should develop evaluation metrics that take different levels of error severity into account (e.g., by weighing the different kinds of errors). After all, this would probably improve the correlation between automatic measures and human judgments. But here we might ask ourselves: *what is quality, really?* Is it some abstract construct that we aim to approach through human ratings? Or do we want to model human responses to textual output? If the former, then our study only shows that human ratings are biased against specific kinds of errors, and we may not want to depend on human ratings too much. If the latter, then weighing different kinds of errors might be a good first step.

6.7 Limitations

We identify three main limitations of our work:

1. Assumptions about gender. Larson (2017) discuss the implications of using gender as a variable in NLP research. In light of their study, we should note that we are manipulating gender as a binary variable; protagonists are either described as a man/woman or as a boy/girl. This is a simplification for the sake of our experiment, to see how people respond to identification errors where *people who are perceived as male* are described as female and vice versa. Because the authors manually identified the gender of the protagonists, these gender labels could be different from the protagonists' actual gender identity.

Whether image descriptions should contain references to gender is a subject of debate. On the one hand, blind or visually impaired users indicate that they would like to see them (Stangl et al., 2020), but on the other hand, gender is notoriously difficult to detect (Buolamwini and Gebru, 2018), and misgendering individuals can be harmful to users (Keyes, 2018). For this reason, Google decided to no longer use gender labels for its image recognition services (Ghosh, 2020).

2. Variation within categories. A fundamental problem with our current line of research is that textual descriptions can be wrong in many ways. As noted in Section 4.2, there is likely also variation within each error category based on the degree of 'wrongness'. Presumably, *orange→red* is less wrong than *orange→blue*. Similarly, *baby→toddler* is probably better than *baby→adult*. This complicates the comparison of different error categories. We aimed to minimize this issue by generating clear-cut mistakes in each category. Still, some variation may remain. In future work, we will investigate this issue further by explicitly targeting within-category variation.

3. No visually impaired end-users included. Finally, we caution that our results only hold for participants with regular vision, and not necessarily for blind or visually impaired users, for whom image description technology is currently being developed. Although there have been some studies on blind or visually impaired users' experiences with this technology (Zhao et al., 2017; Wu et al., 2017), more work is needed to understand the impact of erroneous output on these users. A major challenge in this area is that blind or visually impaired users are not able to determine whether a given image description is correct or not. This means that future work should investigate the impact of different kinds of errors using other means, such as (contextual) interviews or focus groups.

7 Conclusion

We carried out a tightly controlled study, comparing minimal pairs of image descriptions with different types of errors. Our results reveal big differences in perceived quality between these descriptions. Moreover, we even found preliminary evidence that there are also differences *within* error categories. Our results show that we need to take a closer look at the determinants of description quality, and take seriously the idea of different levels of importance for different aspects of an image. On a broader level, gradations in error severity are probably not limited to image descriptions alone. We encourage researchers in NLG to take a closer look at common output errors in their domain, and to consider the different impact that each of those errors may have.

8 Acknowledgments

This study was approved by the Research Ethics and Data Management Committee at the Tilburg School of Humanities and Digital sciences, Tilburg University (reference number: 2019.40 - amended September 26, 2019). The experimental design and main data are from the Master's thesis of Wei-Ting Lu, supervised by Emiel van Miltenburg.

References

Ali Abdolrahmani, William Easley, Michele Williams, Stacy Branham, and Amy Hurst. 2017. Embracing errors: Examining how context of use impacts blind individuals' acceptance of navigation aid errors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 4158–4169, New York, NY, USA. Association for Computing Machinery.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.

Francis J Anscombe. 1973. Graphs in statistical analysis. *The American Statistician*, 27(1):17–21.

Adriana Baltaretu, Emiel J. Krahmer, Carel van Wijk, and Alfons Maes. 2016. Talking about relations: Factors influencing the production of relational descriptions. *Frontiers in Psychology*, 7:103.

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language*, pages 32–68.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

Jean Carletta and Christopher S. Mellish. 1996. Risk-taking and recovery in task-oriented dialogue. *Journal of Pragmatics*, 26(1):71 – 107.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

Kees van Deemter and Ehud Reiter. 2018. Lying and computational linguistics. In Jörg Meibauer, editor, *The Oxford Handbook of Lying*. Oxford University Press.

Lee R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

Shona Ghosh. 2020. Google ai will no longer use gender labels like 'woman' or 'man' on images of people to avoid bias. *Business Insider*. Retrieved: 23 July 2020.

Yvette Graham, George Awad, and Alan Smeaton. 2018. Evaluation of automatic video captioning using direct assessment. *PLOS ONE*, 13(9):1–20.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz Grand Challenge : Answering Visual Questions from Blind People. In *Proceedings of the 2018 Conference on Computer Vision and Pattern Recognition (CVPR'18)*, pages 3608–3617, Salt Lake City, Utah. Computer Vision Foundation.

Nikolaus P Himmelmann and Beatrice Primus. 2015. Prominence beyond prosody: A first approximation. In *pS-prominenceS: Prominences in Linguistics. Proceedings of the International Conference*, pages 38–58. Disucom Press, University of Tuscia Viterbo.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28, Berlin, Germany. Association for Computational Linguistics.

Wenyu Huang. 2017. Chinese classifier categorizations and the application to second language acquisition. Master's thesis, Leiden University.

Laurent Itti and Christof Koch. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489 – 1506.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Xirong Li, Xiaoxu Wang, Chaoxi Xu, Weiyu Lan, Qijie Wei, Gang Yang, and Jieping Xu. 2018. COCO-CN for cross-lingual image tagging, captioning and retrieval. *CoRR*, abs/1805.08661.

Hunter T. Lockwood and Monica Macaulay. 2012. Prominence hierarchies. *Language and Linguistics Compass*, 6(7):431–446.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Emiel van Miltenburg and Desmond Elliott. 2017. Room for improvement in automatic image description: an error analysis. *CoRR*, abs/1704.04198.

Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, 4:21.

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. Association for Computational Linguistics.

Rohit Parikh. 1994. Vagueness and utility: The semantics of common nouns. *Linguistics and Philosophy*, 17(6):521–535.

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

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Stanley Smith Stevens. 1975. *Psychophysics: Introduction to its Perceptual, Neural, and Social Prospects*. New York: John Wiley.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Kees Van Deemter. 2012. *Not exactly: In praise of vagueness*. Oxford University Press.

Jorrig Vogels, Emiel Krahmer, and Alfons Maes. 2013. When a stone tries to climb up a slope: The interplay between lexical and perceptual animacy in referential choices. *Frontiers in Psychology*, 4:154.

Timothy Williamson. 2002. *Vagueness*. Routledge.

Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1180–1192, New York, NY, USA. Association for Computing Machinery.

Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. 2013. Studying relationships between human gaze, description, and computer vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'13)*, pages 739–746.

Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

A Prompt

Figure 4 (repeated from Figure 3) provides an example item. Before each item, there was an instructional text, and the sliders were accompanied by additional instructions. These are provided in Table 4 below, both in the original Chinese, and translated into English.

B Replacements at a glance

Table 5 presents all replacements made in our experiment, with counts for how often each replacement was made to generate an erroneous description.

Instructions	請仔細閱讀以下圖片及文字，並移動滑塊以評估自動生成圖像描述的品質 'Please read the following pictures and text carefully, and move the slider to evaluate the quality of the automatically generated image description'
Correct cue	正確描述 'correct description'
Erroneous cue	自動生成描述 'automatically generated description'
Slider text	請移動滑塊以評估自動生成描述的品質 'Please move the slider to evaluate the quality of the automatically generated description'

Table 4: Instructions for the participants.

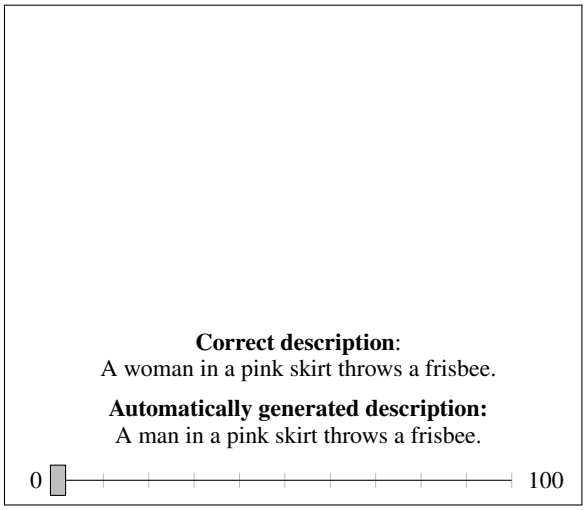

Figure 4: Example item, with the picture (137767 in MS COCO), the reference description, the erroneous description, and a slider to indicate the description quality. Descriptions have been translated and edited for ease of presentation. Original picture taken by Mike LaCon (CC BY-SA 2.0).

Category	Original	Replacement	Count
Gender	man	woman	3
	woman	man	2
	girl	boy	1
	boy	girl	1
Age	man	boy	3
	woman	girl	2
	girl	woman	1
	boy	man	1
Clothing type	shirt	coat	3
	dress	suit	1
	shorts	trousers	1
	suit	swimsuit	1
	skirt	pants	1
Clothing color	black	pink	1
	yellow	purple	1
	black	white	1
	black	red	1
	blue	orange	1
	gray	yellow	1
	pink	blue	1

Table 5: Replacements made in our experiment

C Descriptions

Tables 6-12 (see next two pages) are all the descriptions we used for the images. Images themselves are not provided here, but instead we provide the image ID from the MS COCO dataset. See the images here: `https://cocodataset.org/#explore?id=ID` (replace ID with the actual ID).

Correct	一位	男童	穿著	黑色	上衣	在	棒球場		投球
	A	boy	wear	black	shirt	on	baseball field		pitch
Translation	A boy in a black shirt pitches at the baseball field.								
Gender error	一位	**女童**	穿著	黑色	上衣	在	棒球場		投球
	A	**girl**	wear	black	shirt	on	baseball field		pitch
Age error	一位	男**人**	穿著	黑色	上衣	在	棒球場		投球
	A	**man**	wear	black	shirt	on	baseball field		pitch
Clothing type error	一位	男童	穿著	黑色	**大衣**	在	棒球場		投球
	A	boy	wear	black	**coat**	on	baseball field		pitch
Clothing color error	一位	男童	穿著	**粉色**	上衣	在	棒球場		投球
	A	boy	wear	**pink**	shirt	on	baseball field		pitch

Table 6: Image 320785 from MS COCO

Correct	一位	男人	穿著	黃色	上衣	在	網球場	打	網球
	A	man	wear	yellow	shirt	on	tennis court	play	tennis
Translation	‘A man in a yellow shirt plays tennis on the tennis court.’								
Gender error	一位	**女人**	穿著	黃色	上衣	在	網球場	打	網球
	A	**woman**	wear	yellow	shirt	on	tennis court	play	tennis
Age error	一位	男**孩**	穿著	黃色	上衣	在	網球場	打	網球
	A	**boy**	wear	yellow	shirt	on	tennis court	play	tennis
Clothing type error	一位	男人	穿著	黃色	**大衣**	在	網球場	打	網球
	A	man	wear	yellow	**coat**	on	tennis court	play	tennis
Clothing color error	一位	男人	穿著	**紫色**	上衣	在	網球場	打	網球
	A	man	wear	**purple**	shirt	on	tennis court	play	tennis

Table 7: Image 344149 from MS COCO

Correct	一位	女童	穿著	藍色	洋裝	在	衝浪板	旁	站著
	A	girl	wear	blue	dress	on	surfboard	by	standing
Translation	‘A girl in a blue dress stands by the water park								
Gender error	一位	**男童**	穿著	藍色	洋裝	在	衝浪板	旁	站著
	A	**boy**	wear	blue	dress	on	surfboard	by	standing
Age error	一位	女**人**	穿著	藍色	洋裝	在	衝浪板	旁	站著
	A	**woman**	wear	blue	dress	on	surfboard	by	standing
Clothing type error	一位	女童	穿著	藍色	**西裝**	在	衝浪板	旁	站著
	A	girl	wear	blue	**suit**	on	surfboard	by	standing
Clothing color error	一位	女童	穿著	**橘色**	洋裝	在	衝浪板	旁	站著
	A	girl	wear	**orange**	dress	on	surfboard	by	standing

Table 8: Image 372182 from MS COCO

Correct	一位	男人	穿著	灰色	上衣	在	街道	站著
	A	man	wear	gray	shirt	on	street	standing
Translation	A man in a gray shirt stands on the street.							
Gender error	一位	**女人**	穿著	灰色	上衣	在	街道	站著
	A	**woman**	wear	gray	shirt	on	street	standing
Age error	一位	男**童**	穿著	灰色	上衣	在	街道	站著
	A	**boy**	wear	gray	shirt	on	street	standing
Clothing type error	一位	男人	穿著	灰色	**大衣**	在	街道	站著
	A	man	wear	gray	**coat**	on	street	standing
Clothing color error	一位	男人	穿著	**黃色**	上衣	在	街道	站著
	A	man	wear	**yellow**	shirt	on	street	standing

Table 9: Image 141759 from MS COCO

Correct	一位	女人	穿著	粉色	裙子	在	草地	丟	飛盤
	A	woman	wear	pink	skirt	on	grass	throw	frisbee
Translation	‘A woman in a pink skirt throws a frisbee on the grass.’								
Gender error	一位	**男**人	穿著	粉色	裙子	在	草地	丟	飛盤
	A	**man**	wear	pink	skirt	on	grass	throw	frisbee
Age error	一位	女**童**	穿著	粉色	裙子	在	草地	丟	飛盤
	A	**girl**	wear	pink	skirt	on	grass	throw	frisbee
Clothing type error	一位	女人	穿著	粉色	**褲**子	在	草地	丟	飛盤
	A	woman	wear	pink	**pants**	on	grass	throw	frisbee
Clothing color error	一位	女人	穿著	**藍**色	裙子	在	草地	丟	飛盤
	A	woman	wear	**blue**	skirt	on	grass	throw	frisbee

Table 10: Image 137767 from MS COCO

Correct	一位	男人	穿著	黑色	西裝	在	廁所	自拍	
	A	man	wear	black	suit	in	toilet	selfie	
Translation	‘A man in a black suit takes a selfie in the toilet’								
Gender error	一位	**女**人	穿著	黑色	西裝	在	廁所	自拍	
	A	**woman**	wear	black	suit	on	toilet	selfie	
Age error	一位	男**童**	穿著	黑色	西裝	在	廁所	自拍	
	A	**boy**	wear	black	suit	on	toilet	selfie	
Clothing type error	一位	男人	穿著	黑色	**泳**裝	在	廁所	自拍	
	A	man	wear	black	**swimsuit**	on	toilet	selfie	
Clothing color error	一位	男人	穿著	**白**色	西裝	在	廁所	自拍	
	A	man	wear	**white**	suit	on	toilet	selfie	

Table 11: Image 218368 from MS COCO

Correct	一位	女人	穿著	黑色	短褲	在	網球場	打	網球
	A	woman	wear	black	shorts	on	tennis court	play	tennis
Translation	A woman in black shorts plays tennis on the tennis court.								
Gender error	一位	**男**人	穿著	黑色	短褲	在	網球場	打	網球
	A	**man**	wear	black	shorts	on	tennis court	play	tennis
Age error	一位	女**童**	穿著	黑色	短褲	在	網球場	打	網球
	A	**girl**	wear	black	shorts	on	tennis court	play	tennis
Clothing type error	一位	女人	穿著	黑色	**長**褲	在	網球場	打	網球
	A	woman	wear	black	**trousers**	on	tennis court	play	tennis
Clothing color error	一位	女人	穿著	**紅**色	短褲	在	網球場	打	網球
	A	woman	wear	**red**	shorts	on	tennis court	play	tennis

Table 12: Image 35948 from MS COCO