Gradations of error severity in Automatic Image Descriptions in English

Anastasios Heracleous

MSc in Artificial Intelligence

Supervisor: Guanyi Chen
Second Examiner: Yupei Du

Contents

Abstract

There is a growing interest around the evaluation of automatic image description systems by evaluation metrics and their poor correlation with human evaluation scores for automatically generated text. Following a research carried by Miltenburg et al. (2020) we performed an experiment on English speakers to validate if different kinds of errors in image descriptions, elicit different evaluation scores. We performed an experiment in two parts. The first part contained human text descriptions paired with descriptions that we have manipulated to contain errors and the second part had he same structure but with pictures as well. Participants evaluated the quality of the manipulated descriptions compared to the human descriptions for the first part and compared to the human descriptions and the pictures in the second part. Our results show that the severity of different kinds of errors is perceived differently by humans which give different evaluation scores to each error type according either solely to the text they have read or the picture they have seen. Evaluation metrics failed to capture these differences, thus their poor correlation with human judgements. In an attempt to understand why different errors are seen as more or less severe, our work provides the foundations of the influence an image plays in the way humans evaluate different kinds of errors.

1 Introduction

Recent technological advancements in Natural Language Processing (NLP) and Computer Vision (CV) have made it possible to create programs that can generate descriptions for images which are called Image Captioning [1]. But these automatically generated descriptions might contain some errors. Therefore in order to assess the quality of such a program you need to determine a form of evaluation. One might say to let humans just decide if a description produce by the system is of good or bad quality. Humans are able to relatively easily describe the environments they are in. Given an image, it is natural for a human to describe an immense amount of details about this image with a quick glance [2]. Gathering humans to evaluate automatically generated descriptions might not seem practical when you have a million descriptions to evaluate. Recent years have seen a growing popularity in automatic evaluation as it is cheaper and faster to run than human evaluation. The use of such metrics is only sensible if they show sufficient correlation with human judgements. Sensible in a way that our aim is to replace human evaluators with automatic evaluation metrics. If we know that these automatic evaluation metrics do not correlate with human judgments then we cannot really replace the human factor in the evaluation of automatic image descriptions. The strong correlation between those two is rarely the case as shown by various studies in Natural Language Generation (NLG) [3], [4], [5], [6]. While many researchers prove that several metrics have poor correlation with human judgements, only recently we have seen an interest in explaining why there is such poor correlation. An experiment run by Miltenburg et al. 2020 on Chinese speakers focused on the evaluation of image descriptions. In their experiment, they systematically manipulated image descriptions to contain different kinds of errors. Their aim was to check if different kinds of errors elicit different evaluation scores, which is what their results showed. Evaluation metrics that are based in textual similarity are unable to capture such differences and every error is treated in the same way when evaluating automatic image descriptions. This might partially give away one of the reasons we see such poor correlation between human judgements and automatic evaluation metrics. Their research is also the basis of this paper where we want to see if their findings

generalise with English speakers and also get new insights on the perspective humans have when evaluating errors in image descriptions. Automatic image description systems make different errors and these errors are likely to be of different importance. Consider Figure 1, which shows multiple human reference descriptions and a description generated by Aracil's model 3 [7].

This system makes three different mistakes, which are shown separately in Example 1. We refer to these mistakes as an object error (1b), age error (1c) and activity error (1d).

Intuitively, the different errors made by the system are not equally severe. By looking at the picture, our intuition is that object error is more severe than the activity error. It is not that clear in the picture that the girl is not sitting. In section 4 we will discuss the influence a picture makes when deciding on the severity of an error, where we posit our hypotheses. This paper provides evidence that there are differences in perceived error severity between different kinds of errors in image descriptions for English speakers. Most metrics wrongly assume that there is no difference between different kinds of mistakes. Our results showed that even when based only on text, humans assign different evaluation scores for different kinds of errors. When there is an image present, the results showed again this differentiation but the errors received even lower scores. Thus the correlation between humans and automatic evaluation metrics will not improve unless we try different approaches.

The structure of this paper will go as follows: Section 2 provides related work in the natural language generation and image captioning research areas. An outline of several evaluation metrics that have been developed throughout the years and the problem with them. Section 3 posits our research questions and the hypotheses behind them. Section 4 gives a thoroughly description of the experiment. Section 5 is dedicated to show the results obtained. Finally, section 6 and 7 presents our discussion of the results and the conclusions about this project while we give some possible future work directions.

Human:

- little girl blowing out a candle on a fancy dessert.
- a little girl leaning towards a small cake.
- an Asian child looking at cake with a candle on it, surrounded by other people.

System:

- man is sitting at table with pizza

Figure 1: Image 376295 from the MS COCO dataset, with human descriptions from the COCO-CN corpus, and an automatically generated description from Aracil's model 3 [7]. Errors in the automatically generated description are highlighted in red

(1) Gold standard (a) and errors (b-d) from Figure 1.

a. A girl leaning at the table with small cake

b. A girl leaning at the table with pizza

c. A man leaning at the table with small cake

d. A girl sitting at the table with a small cake

2 Related Work

2.1 Natural Language Generation and Image Captioning

NLG is the subfield of artificial intelligence and computational linguistics that is concerned with the construction of computer systems that can produce understandable texts in English or other human languages from some underlying non-linguistic representation of information[8]. Taking a text as an input and producing automatically new,coherent text as output falls under the description of text-to-text generation and is a huge aspect of NLG. Text-to-text generation can generate complete sentences and form essays from just some meaningful words [9], [10], [11]. Machine translation is another well known example of text-to-text generation [12]. Another instance of NLG is data-to-text generation. It generates natural language descriptions conditioned on structured input like datasets, tables or even images [13]. Image captioning which is a paradigm case of data-to-text generation takes as input images and outputs descriptions of images. More precisely it is a task where you analyse a picture and generate a description that expresses the most salient aspects of it. There are different methods for image captioning that chronologically developed and try to find the optimal way to description generation. The earliest approaches we could see are the Template-based approaches where different meaning tuples refer to different aspects of the image ([14], [15], [16]). At first you detect objects, actions and attributes and then the blank spaces in the templates are filled. Li et al. [17] extract phrases related to the detected objects and attributes. A downside of template-based approaches is that the templates are pre-defined so you cannot generate variable-length captions. Later on, it became popular the research on retrieval-based approaches. We can have a set of existing captions in a visual or multi-modal space and retrieve them whenever we want. These methods, first find images that are visually similar along with their captions from the training dataset, the so-called candidate captions. When you enter a new image as input to the system, it will search the caption pool to find the most relevant caption. Gong et al. [18] proposed such a method that generates descriptions from the multi-modal space and also uses information retrieval techniques to find the

most suitable description to return. Similar approaches were also proposed by Hodosh et al. [19], Hodosh, M., and Hockenmaier, J. [20], Ordonez, V. et al. [21] and Sun et al.[22]. Even though the sentences that are retrieved are well-formed human-written sentences or phrases, these are only specific to already known images to the system and cannot adapt to new inputs, while also in some cases the generated description ends up to be completely irellevant to the image contents [23]. Retrieval-based and template-based are adopted mainly in early work and due to great progress a new method involving neural networks became the current trend on image captioning [24], [25]. The deep-learning image captioning methods first analyse the visual input and then a language model to generate image-related captions. Kiros et al., [26] used a method adopting Convolutional Neural Network to extract the image features in generating image captions. Kiros R., Salakhutdinov R and Zemel S. [27] introduced a general encoder-decoder framework that will allows a sentence output to be generated word by word given an image input. The encoder creates the joint multimodal space which is used for the ranking of images and descriptions. The decoder uses that multimodal representations to generate descriptions. It is a combination of both Long–Short Term Memory (LSTM) Recurrent Neural Network to encode any text [28] and a deep convolutional network to encode the image features. Encoded visual data is projected into an embedding space extended with the LSTM hidden states. The embedding space allows the mapping of visual and textual features to the same latent space, so both of them can be processed together using machine learning [29]. Kiros et al uses a structure-content neural language model in the embedding space to decode decode visual features conditioned on context word feature vectors, allowing for sentence generation word by word [30]. Similar approaches were also proposed by Karpathy et al. [31] and Yagcioglu et al. [32]. There are different approaches developed throughout the years as discussed above and quite recently the adoption of Reinforcement learning techniques is being researched on [33], [34], [35].

2.2 Evaluation Metrics

One of the very first metrics to compare human descriptions with automatically generated descriptions and their accuracy was the BLEU metric [36]. It is used to calculate the co-occurrence frequency of two sentences (the candidate sentence and the reference sentence) based on the weighted average of matched n-gram phrases. The problems that arose with this approach is that it ignore the meaning of words, i.e. if some words have the same meaning but are not identical then it would consider them as wrong. Also, words that occur from the same lemma, for example "play" and "playing" were not consider a match from the precision so the score remained low. Further, in a short document or sentence, there is a high probability of obtaining zero tri-gram or 4-gram precision, which makes the overall BLEU score equal zero due to the use of geometric mean [37]. As a result, this metric cannot achieve high correlation with human evaluations scores for automatically generated text.

METEOR is a precision and recall based method used to evaluate machine translation [38]. It performs a unigram match (word to word) between the candidate sentence and the human-written reference sentences and then computes a score based on that. It came to resolve the weaknesses of the BLEU metric mentioned before relating to the low scores due to zero tri-gram or 4-gram precision and it has a high correlation with human judgment not only at the entire collection but also at the sentence and segment level.

ROUGE is a recall-oriented metric designed for summarization [39]. It is a set of metrics for evaluating automatic summarization of long texts consisting of multiple sentences or paragraphs. ROUGE includes the mean or median score from individual output text, which allows for a significance test of differences in system-level rouge scores, while this is restricted in BLEU [40].However, ROUGE has problems, among others, in evaluating multi-document text summaries [41].

CIDEr is an automatic metric for measuring the similarity of a generated sentence against a set of human-written sentences using a consensus-based protocol [42]. An interesting intuition of this metric is that it tackles one of the disadvantages of the BLEU metric which treated all words on the match the same, and actually treats some words as more

important than others using Term Frequency Inverse Document Frequency (TF-IDF) weight calculation for each n-gram. CIDEr tends to favor more descriptive captions (likely due to preference for rarer n-grams) but this are not always the most accurate descriptions of an image.

SPICE measures how well caption models recover objects, attributes and relations [43]. When it comes to image captioning, instead of focusing on n-gram similarity, SPICE gives more importance to the semantic propositions implied by the text. There is room for improvement as currently the performance with large amount of reference caption is not very good. A major drawback of SPICE is that it ignores the fluency of the generated captions [44] Liu et al. [45] introduced a new caption evaluation metric that is a good choice by human raters. It is developed through a combination of SPICE and CIDEr, and termed as SPIDEr. It uses a policy gradient method to optimize the metrics.

BERTScore is metric that can evaluate various language generation tasks including image captioning [46]. It exploits pre-trained BERT embeddings [47] to represent and match the tokens in the reference and candidate sentences via cosine similarity. The best matching token pairs are used for computing precision, recall, and F1-score. However, while its performance on NLP tasks set a new state of the art in general, studies of specific syntactic and semantic phenomena have shown where BERT's performance deviates from that of humans more generally [48].

2.3 The problem of Evaluation Metrics

Several papers reported that automatic evaluation metrics do not correlate with human evaluations [3], [49], [50] and as new evaluation metrics are developed in an attempt to achieve better correlation, there are no studies to explain what are the reasons we see such a poor correlation [51]. In their paper, Miltenburg et al. (2020), proposed one of the reasons as to why we see such poor correlations by focusing on the evaluation of automatic image description systems. Image description systems make different kinds of errors and their intuition is that some errors are more severe than others. For example, given an image showing a 29 year old woman holding a cake, the hypothesis is that the object

error is considered as more severe than the age error. Age is a more vague attribute than object category. Miltenburg et al. (2020) carried an experiment on Chinese speaking participants to test how they rate erroneous descriptions when compared to correct ones with a scale of 0(worst) to 100(best). The main finding was that different error types vary in severity and suggests that people attach different levels of importance to different aspects of an image description and this can also be seen by tables 1,2 and 3. Moreover the findings showed that clothing color errors are significantly worse than clothing type errors which came as a surprise, as the expectation was the opposite result. The results found reveal big differences in perceived quality of image descriptions with different types of errors but also give some evidence for differences within error categories.

For example changing male to female, versus female to male. The results showed a significant effect of error directionality for age and more specifically a reduction in description quality when changing the label from old to young. This was believed to be the case because of Chinese culture and politeness but if this significant effect of error directionality still exists with participants of other ethnicity then the interpretation of the result would be different. Through their findings it was made clear that more work is needed as 2 out of 3 hypotheses were disproved and still it is not clear as to why there are these differences in severity. The results gave a starting point on why these differences in severity arise and with the addition of more error types the scene can become clearer. Evaluation metrics try to compare this automatic image captioning with human judgements and the results are not as promising. The metrics fail in several cases to distinguish between human-written and machine generated captions. One of the many reasons of the poor performance of the automatic evaluation metrics and a serious problem in nlg output is approached by Miltenburg et al., [52]. Different kinds of Natural Language Generation systems make different kinds of errors but unfortunately there is severe under reporting of them. Out of 111 papers collected from related NLG conferences, it was found that only 5 papers included error analysis, thus in only 5 systems we could know what is going wrong and possible ways to fix it.

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 [51].

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 a = 0.05 [51].

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 [51].

The paper by Miltenburg et al. also provides recommendations for error identification, analysis and reporting which comes close with the general idea of this paper to categorise different types of errors in automatic image descriptions and treat each category differently.

3 Research Questions and Hypotheses

Looking at possible explanations on why there is such a poor correlation between human judgements and evaluation metrics we come across on a significant difference in the way error types are treated by humans and by metrics. The experiment carried by Miltenburg et al. (2020), was based on Chinese speaking participants and the results proved that different kinds of errors elicit significantly different evaluation scores. The next step is to check if these results generalise in other languages as well but by keeping in mind cross-linguistic differences. The self-denigration Maxim exists in Chinese culture but not in English culture and might influence the comprehension of an error [53]. To be more specific, Chinese crucially differs from English in that the word 'girl' cannot be used to refer to an adult woman, 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 by English native speakers. Therefore the main research question of this study is:

RQ1: "Do native English speakers react in a similar way as native Chinese speakers given error categories within image descriptions?" To test this research question we performed the same experiment as Miltenburg et al. (2020) did but everything was translated to English and the participants were either native speakers or at an advanced professional proficiency level.

We came to the conclusion that there are big differences in perceived quality between different types of errors but we still need to figure out why different errors are seen as more or less severe. A possible explanation to that is the influence we get from images. Therefore the second research question of this study is:

RQ2: Will the presence of an image make any impact on the perspective of the participants on their description quality evaluation?

In order to explore this research question, we are going to give to participants the same survey but twice. The first time they are going to come across with a survey that contains only text and no pictures, while the second survey will contain pictures as well. In this way we hope to test if the levels given at each error type category in the first survey will

be the same in the second survey when a picture is present.

H1-H3 are the hypotheses stated by Miltenburg et al. (2020) at their experiment and we want to test if their results would generalise with English speakers. If we get the same responses, i.e. H1 and H3 disproved and H2, confirmed then we would obtain a strong correlation between the Chinese participants in Miltenburg's experiment and the English participants in this experiment.

In the case that H4 will be confirmed this will add to the theory that the degree of vagueness affects our perception. Clothing types can be a huge amount of different categories with many of them be similar to each other thus not making it a salient feature of an image [54]. For example, A shirt can be similar to a polo shirt as both have buttons and such an error is likely to not be consider as more severe when a picture is present in contrast when there is no picture.

The framework discussed by Itti et al. (2001) suggests that subjects selectively direct attention to objects in a scene using both bottom-up, image-based saliency cues and top-down, task-dependent cues. Some stimuli are intrinsically conspicuous or salient in a given context [55]. Presenting a clothing color error with an image accompanying it, e.g. from black to red, to a human can be considered a salient one and immediately pops out from the visual scene. In the case that H5 will be confirmed, this would support the above findings by Itti et al. (2001) and the weakness of the eye movements into color contrast. Further, clothing color is a salient feature of an image while also the only color mentioned in the description, and salient features of an image are always described and given emphasis to when asked to describe them [54].

In the case of H6, gender related errors are not debatable. It is either male or female the acceptable answers to describe a person in a picture thus limiting the degree of vagueness. When describing a picture, the human eye tends to search for the salient features of the picture to mention [54]. In our experiment, there is a clear view of the person's face and body in every picture, thus there is no debate as to what gender the person in the picture is.

H1: The perceived quality of descriptions with people-related errors is lower than the

perceived quality of descriptions with clothing-related errors

H2: The perceived quality of descriptions with clothing color error is higher than the perceived quality of descriptions with clothing type error.

H3: The perceived quality of descriptions with age-related errors is higher than the perceived quality of descriptions with any other error type from the given ones.

H4: The perceived quality of descriptions with clothing type-related errors when there is no picture will be similar or the same than when there is.

H5: The perceived quality of descriptions with clothing color-related errors when there is a picture will be lower than when there is not.

H6: The perceived quality of descriptions with gender-related errors when there is no picture will be higher than when there is.

4 Experiment

4.1 Overview

The following experiment was designed to test the proposed hypotheses. In this experiment, the evaluation scores for the quality of automatically generated text descriptions were collected. There were no criteria or guidance on how to evaluate the quality of the text presented to the participants and the scores were solely based on the participants' opinion. The experiment was divided into two different parts, in order to identify and test the hypotheses above. The first part of the experiment focused on the text description alone and the scores were based on what the user reads while the second part included a picture for each question. Both parts of the experiment took place on Qualtrics during the month of May 2022. The stimuli used was the same for both parts and will be explained thoroughly below along with the procedure.

4.2 Participants

A total number of 50 people were recruited to participate in this experiment (24 female, 26 male; 24 native, 24 fluent speakers of English) and were recruited via the researcher's social media channels. Every participant received a university education.

4.3 Materials

The materials used for this experiment followed the decisions made by Miltenburg et al. (2020) and were overall the same. The image selection was made from MC COCO and the same 7 images were picked. For each image there was a need for a manual construction of 4 descriptions which were translations from the original paper as those were in the Chinese language. Each description had one error, resulting in 28 image-description pairs. Figure 2 shows an example image with the reference description, and four erroneous descriptions. Regarding the image selection the following criteria were met when selecting the images from the MS COCO dataset.

Correct Translation: A boy in a black shirt pitches at the baseball field

Gender Error: A girl in a black shirt pitches at the baseball field

Age Error: A man in a black shirt pitches at the baseball field

Clothing type error: A boy in a black coat pitches at the baseball field

Clothing color error: A boy in a pink shirt pitches at the baseball field

Figure 2: Correct reference description, along with systematically manipulated descriptions for image 320785 from the MS COCO dataset. Each erroneous description contains only one word that has been altered compared to the original one.

1. Colorful 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

The aim is to avoid additional variance in the experience by eliminating error ambiguity. For example, if the boy in Figure 2 was wearing black pants as well, then the clothing type error could be resolved in two ways: change coat to shirt or change coat to pants Regarding the descriptions, as it was mentioned above we divided them into four error type categories which all relate to the PEOPLE main category following the annotation scheme developed in van Miltenburg and Elliott (2017) [56] and one category that consists of the correct descriptions. Table 4 shows the 5 categories and the number of descriptions included in each category. This categorization gave us the possibility to run further analysis on the stimuli collected.

Category	Count
Correct	7
Age	7
Gender	7
Clothing type	7
Clothing color	7

Table 4: The 5 categories and the number of descriptions included in the stimuli for each category

The category "correct" includes descriptions that did not contain any error. The category "Age" includes descriptions containing a mistake in the age; for example, girl versus woman. The category "Gender" includes descriptions containing a mistake in the gender; for example, a girl versus a boy. The category "Clothing type" includes descriptions containing a mistake in the piece of clothing worn by the human subject in the image; for example shirt versus coat. Finally, the category "Clothing color" includes descriptions containing a mistake in the color of a piece of clothing worn by the human subject in the image; for example black shirt versus pink shirt. In order to avoid bias, we divided each description into pairs (correct description + one of the error categories for each images) and assigned them numbers. We had a total of 28 descriptions so we used a random generator for numbers between 1-28 to determine the order that each pair will appear on the survey. When a number that was already included in the survey appeared again, was ignored until all pairs were entered in the survey, in the order that were taken from the random generator. Due to the randomness factor, there was a case where the same picture appeared in two consecutive questions. That was not considered an issue, therefore no measures were taken to rearrange the order. Again the aim was to avoid error ambiguity in descriptions. For example, suppose that the boy in Figure 2 were erroneously referred to as wearing black pants. We could resolve this issue in two ways: (1) resolve the clothing: black shirt, (2) resolve the color: white pants. It is not clear which error type is applicable and these kinds of ambiguities make it impossible to determine the impact of individual error types. Therefore the descriptions provided have only a single fix with the lowest edit distance. Further, there was a high risk for additional variance in the colour error type. If there is a mistake with a hue of a color,

for example red - orange that might be considered less severe than the error: red - black. So, the descriptions produced were clear cur examples for each error category.

4.4 Design

The experiment was implemented in Qualtrics, and followed a within-subjects design, where each participant was exposed to all 56 stimuli (i.e.,2 surveys, 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. The erroneous description was always presented in the context of the image and the correct reference description.

4.5 Procedure

As it was mentioned before, the participants were recruited through different social media channels where the Qualtrics link was privately provided to them. After clicking the link, they were first shown an introductory text with the description of the study (its purpose) and some instructions on how to answer the questions (look at Figure 3)

Pilot study: For the pilot study, the experiment was given to two participants to provide feedback and make sure that it serve its purpose while everything was understandable to people with no background on NLG or any related topics. From the feedback, it was shown that there was need for some rephrasing of the instructions to make them more clear while also asked to provide a small instruction to every question because after a few questions, participants forgot how they were supposed to answer. The two participants had hard time understanding what they were suppose to evaluate, thinking at first that it was expected from them to rate how related the two descriptions (the correct one and the generated one) provided were. As this was not exactly the case, there was a need for a restructure of the instructions, to make it clearer that the participant is suppose to rate how serious is the error made regarding the quality of the description. If we did not change the instructions there was a risk that the participants would end up evaluating something else.

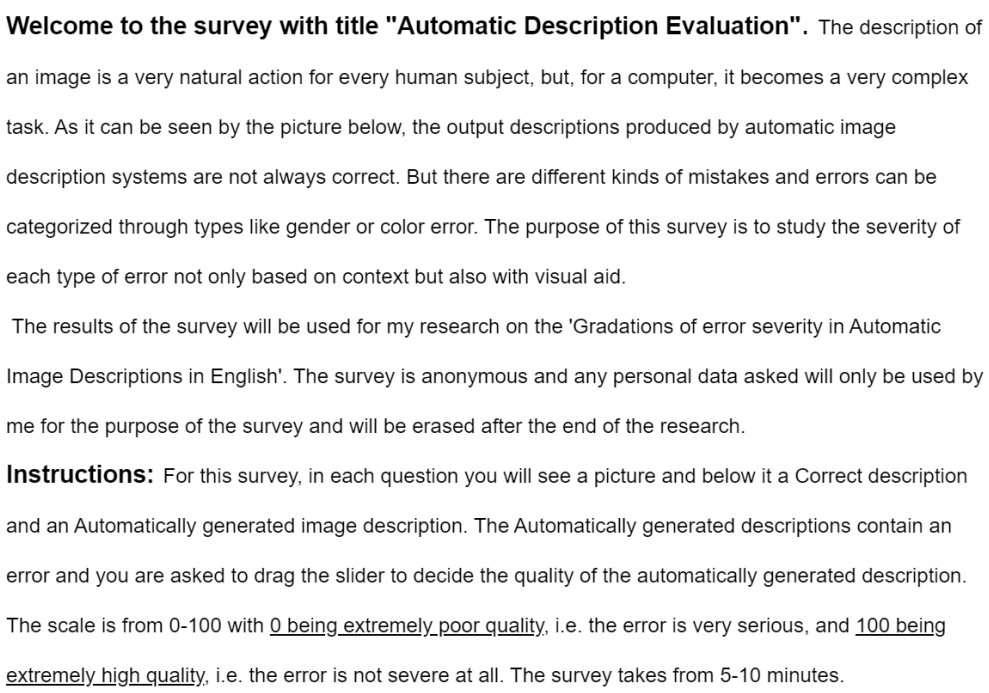

Figure 3: Introductory text of the survey with instructions

For example, if someone thought that they have to compare the two sentences, then they can say that since only one word is different in each question, the scores would be very high in each case. Figure 4 shows how the questions looked before and after receiving the feedback. Taking the sentence of Figure 4 into consideration the success rate of this question would have been 10/11 with the participants first grasp of the instructions. To avoid this risk, the instructions turned out to be like Figure 3, while also in each question of the survey there was a small instruction on the top to remind the participant that it evaluates the quality of the automatically generated description as we can see from Figure 4. The participants in this phase answered all 28 questions where they were asked to indicate the quality of automatically generated description on a slider bar.

Demographic questions: After reading the instructions, participants were asked to answer a few personal questions; their age, gender and English proficiency.

Age and Gender are used further in the analysis section while the purpose of the English proficiency question was to check the validity of the participant.

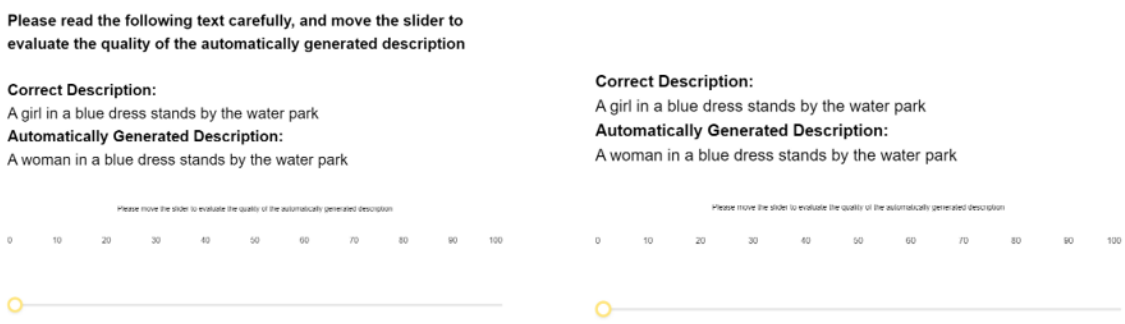

Figure 4: The left picture is how the question looked after the feedback and the right picture is what it looked before

As one of the research questions is to evaluate if the results found in the Chinese experiment held by Miltenburg et al. (2020) would generalise with English participants, there was a condition for participants to speak at a professional level the English language or be a native speaker. When a participant gave a different answer in the English proficiency question rather than the professional or native option, then their records would be considered void and not be used in the analysis phase.

Main Experiment: The main experiment featured the same questions as in the trial phase. Each participant was asked to rate the quality of all 28 stimuli, presented in random order. In order to test RQ2 the experiment was split into two parts. The first part was the 28 questions but without any pictures, an example question can be seen in Figure 4 and the second part was the same 28 questions in the same order but now with pictures, an example question can be seen in Figure 5.

Before running our study, we carried out a pretest to get feedback, and to determine the duration of our experiment (10-15 minutes), to inform the participants before taking part in the study.

Post-Processing: Before the analysis of the results could begin, some processing needed to take place for the preparation of the data. There were five participants that their responses were declared void and had to be removed. Two of these responses were not taken into consideration for the analysis phase because the applicants chose in the English language proficiency question the "elementary level" option. One of the conditions for this experiment was that the participants had at least a professional working

Correct Description:
A girl in a blue dress stands by the water park
Automatically Generated Description:
A woman in a blue dress stands by the water park

Please move the slider to evaluate the quality of the automatically generated description
0 10 20 30 40 50 60 70 80 90 100

Figure 5: An example question from part 2 of the experiment that includes a picture

proficiency of the English language. The other three responses that were rejected was due to the fact that the participants did not complete the survey. Further, we run into a complication with the qualtrics platform. Figure 5 shows an example of a question from the survey. The task was to evaluate the quality of the descriptions. If the participant decided to give a zero value as an answer for some questions then in some cases they did not even touch the slider because as we can see from Figure 5 it is preset at zero. But the qualtrics platform could not record an answer if the user did not touch or drag the slider. From the feedback that we got from every participant after completing the survey, there were several cases where they evaluated several questions with zero. bearing this in mind and having an instruction at each question saying "Please move the slider to evaluate the quality of the automatically generated description", see figure 4 and 5, allowed us to make the assumption that the participant's responses that had empty cells in some questions but not in all of them, were genuine evaluations with scores of zero that they have just not moved the slider and not the case of a skipped question. As a result, when analysing the results we came across with several empty cells that we had to manually replace with zero. Finally, several columns were dropped as were not needed for the analysis.

5 Results

The analysis of results followed several steps. Eventhough, Field (2019) [57] came to disprove the theory behind ANOVA test being robust, for this experiment we decided to go with one-way ANOVA test as it will serve its purpose and determine whether there are any statistically significant differences between the means of the error type groups for each survey. We found that for both survey parts, different error types are indeed judged differently; A repeated measures ANOVA revealed a significant overall effect of error type

$$F(3, 117) = 27.588, p <= 0.05, \eta^2 = 0.199$$

for the first part of the survey and

$$F(3, 123) = 34.798, p <= 0,05, \eta^2 = 0.204$$

for the second part of the survey.

Tables 5 and 6 provide descriptive statistics, showing the different mean scores and their standard deviations for each part of the survey. As we can see, for the first part of the survey where no picture was presented to the participant, errors that are age related received high evaluation scores which means where not considered as serious as the gender related errors which received low scores overall. Similar results where obtained in the second part of the survey where a picture was present in each question. All the categories received lower evaluation scores, with age related errors receiving the highest scores overall and gender related errors receiving the lowest scores overall. All mean scores were lower at the second part compared to the first part. Age-related error mean scores went from 60.8 to 44.3, gender-related error mean scores went from 27.6 to 16.7, clothing type-related error mean scores went from 49.5 to 37.5 and finally clothing color-related mean scores went from 45.7 to 31.0.

Category	Mean	Standard Deviation
Age	60.8	25.8
Gender	27.6	24.5
Clothing type	49.5	23.9
Clothing color	45.7	24.3

Table 5: Descriptive statistics for each of the error categories for the first part of the survey where no pictures were displayed to participants. Mean scores are on a scale from 0–100, where 0 is bad and 100 is good

Category	Mean	Standard Deviation
Age	44.3	21.9
Gender	16.7	18.4
Clothing type	37.5	21.3
Clothing color	31.0	23.7

Table 6: Descriptive statistics for each of the error categories for the second part of the survey where pictures were displayed to participants. Mean scores are on a scale from 0–100, where 0 is bad and 100 is good

The next step of the data analysis consisted of the evaluation of the hypothesis proposed earlier. To do this, 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 Tables 7 and 8 for the first and second part of the survey respectively. The results obtained show some support for H1: Descriptions containing clothing-related errors are significantly better than those with gender-related errors in both parts of the survey. Errors regarding clothing type seem to be roughly on the same footing as age-related errors. The results show no support for H2: we expected that clothing type errors would be worse than clothing color errors, but in fact we found the opposite in both parts. Clothing color-related errors are significantly worse than clothing type errors in the second part of the survey where there was a picture while in the first part, they received again lower scores than the clothing type errors. Looking at tables 5, 6, 7 and 8 we can find support for H3: We expected that the perceived quality of descriptions with age errors would be higher than any other error categories.

Category 1	Category 2	t	df	p-value	Adjusted p-value	Significant?
Age	Clothing color	3.73	39	6.06e-4	0.004	Yes
Age	Clothing type	2.76	39	9e-3	0.053	No
Age	Gender	8.07	39	7.59e-10	0.00000000455	Yes
Clothing color	Clothing type	-1.42	39	1.63e-1	0.978	No
Clothing color	Gender	5.75	39	1.16e-6	0.00000696	Yes
Clothing type	Gender	5.64	39	1.65e-6	0.0000099	Yes

Table 7: Results of multiple paired sample t-tests to compare the means of the scores for the different error categories for the first part of the survey where no pictures were displayed to participants. The table shows both the original p-values and the Bonferroni-adjusted p-values that were used to determine significance at a = 0.05

Category 1	Category 2	t	df	p-value	Adjusted p-value	Significant?
Age	Clothing color	3.98	41	2.71e-4	e-3	Yes
Age	Clothing type	2.20	41	3.3e-2	1.99e-1	No
Age	Gender	9.59	41	4.9e-12	2.94e-11	Yes
Clothing color	Clothing type	-2.91	41	6e-3	3.5e-2	Yes
Clothing color	Gender	5.26	41	4.81e-6	2.89e-5	Yes
Clothing type	Gender	7.19	41	9e-9	5.4e-8	Yes

Table 8: Results of multiple paired sample t-tests to compare the means of the scores for the different error categories for the second part of the survey where pictures were displayed to participants. The table shows both the original p-values and the Bonferroni-adjusted p-values that were used to determine significance at a = 0.05

That seems to be the case as the mean scores of the age-related errors in both parts of the survey are much higher than other categories while when compared with specific error categories age-related errors significantly differed from clothing color-related errors and gender-related errors.

In order to test hypotheses 4-6 we have generated a new table where we compare the mean scores of age, clothing type, clothing color and gender related errors of the first part of the survey with the respective mean scores of the second part of the survey. This information is portrayed by table 9. There is some evidence to support H4: we expected that the perceived quality of descriptions with clothing-type related errors will be similar in both parts of the survey. This can be seen from table 9 where the two categories do not differ significantly. For H5: we expected that the perceived quality of descriptions with clothing color-related errors will be lower when there is a picture rather than when there is not. That is indeed the case as we found that scores for clothing color errors for

Category_second	Category_first	t	df	p-value	Adjusted p-value	Significant?
Age	Age	2.87	41	0.006	0.181	No
Clothing color	Clothing color	-3.30	41	0.002	0.03	Yes
Gender	Gender	-2.28	41	0.028	0.414	No
Clothing type	Clothing type	-2.64	41	0.012	0.176	No

Table 9: Results of multiple paired sample t-tests to compare the means of the scores for the different error categories of the first part of the survey with the respective ones of the second part. The table shows both the original p-values and the Bonferroni-adjusted p-values that were used to determine significance at a = 0.05. Note: Category-second is for mean scores of the second part of the survey and Category-first is for mean scores of the first part of the survey

the second part of the survey are significantly worse than the respective ones for the first part of the survey. Finally for H6: we expected that the perceived quality of descriptions with gender-related errors when there is no picture will be higher than when there is but as the table shows, there is no significant difference between the two categories. Its worth mentioning that even though there is a big drop in the mean score of age-related errors from part 1 to part 2, when running the t-test we found no significant difference for this error type.

We also looked at differences within different error categories. Specifically, we investigated the direction of the errors for two error types: (1) AGE: changing young to old (e.g. boy→man), versus old to young. (2) Gender: changing male to female (e.g. man→woman), versus female to male. Descriptive statistics are provided in Tables 10 and 11. For the first part of the survey where no picture was available we found that the means for both age-related errors and both gender-related errors are similar, and we failed to find a significant effect of error directionality. For Age we got (t(41)= 1.466, p=0.233) and for Gender we got (t(41)= 0.263, p=0.611). That is clearly not the case for the second part where a picture was present to the participants. There is a significant effect of error directionality for age (t(41)= 4.353, p¡ 0.05) and a significant effect of error directionality for gender (t(41)= 8.5, p¡ 0.05). Changing the label from young (e.g. boy) to old (e.g. man) on average leads to a 5-point reduction in description quality (on a scale from 0 to 100). Changing the label from female (e.g. woman) to male (e.g. man) on average leads to a 4-point reduction in description quality (on a scale from 0 to 100).

Category	Direction	Mean	SD
Age	Young to old	61.938	26.960
Age	Old to young	60.380	26.040
Gender	Male to female	27.800	24.543
Gender	Female to male	27.317	25.280

Table 10: Descriptive statistics for subcategories of AGE and GENDER-related errors for the first part of the survey. Higher score means greater perceived quality.

Category	Direction	Mean	SD
Age	Young to old	40.575	25.064
Age	Old to young	45.765	22.754
Gender	Male to female	18.45	20.637
Gender	Female to male	14.45	17.377

Table 11: Descriptive statistics for subcategories of AGE and GENDER-related errors for the second part of the survey. Higher score means greater perceived quality.

6 Discussion

The present thesis was set up to explore the topic of the lack of a strong correlation between the current automatic metrics used for the evaluation of automatic image description systems and human judgments. Following the experiment by Miltenburg et al. (2020) performed on Chinese participants, we wanted to see if the results would generalise for English participants too, thus our first research question. Tables 5 and 6 are the respective ones to Table 1 that was taken by Miltenburg et al. (2020). Although the error type categorization is not the same (in our experiment, gender type errors were classified as the most serious ones in both surveys while that is not the case with Miltenburg's experiment) we can come to the same conclusion that different error types are indeed judged differently. Tables 7 and 8 relate to Table 2 that is taken by Miltenburg et al. (2020). We compared the means of the scores for the different error categories as Miltenburg did but in this case we got different results. In both our surveys, the difference between the means of the scores of gender-related errors with clothing-related errors is significant while in Miltenburg's experiment it was not significant. This might be due to several reasons like the cultural differences between the two countries or the fact that our experiment was performed in a different way. The first part of the survey is based only on text and no picture is present, therefore participants might have judged more harshly the gender-related errors as they are more flagrant, it is either male or female. The second part of the survey that did consists of pictures, came right after they have completed the first part and the questions were in the same order, so the participants might have been influenced by their previous answers of the first part of the survey. But we got the same significance levels with Miltenburg's experiment when comparing the means of the scores of age-related errors with the other three error types. Thus, we can find some evidence that the English participants evaluate descriptions in similar way with Chinese participants. Tables 10 and 11 related to Table 3 that is taken by Miltenburg et al. (2020). We also find the descriptive statistics for subcategories of AGE and GENDER-related errors. Table 10 that relates to the first part of the survey that did not have any pictures, revealed different mean scores with Miltenburg's experiment, i.e. under the category age the mean

scores are much higher and under the category gender the mean scores are much lower in our experiment. We did not find any significant effect of error directionality in both categories. Assuming that this might be due to the fact that there is not picture in our first part of the survey, we do not consider significant these differences in score of Table 10 with Table 3. Table 11 refers to the results drawn by the second part of the survey that did have pictures. We also found a significant effect of error directionality for age as Miltenburg's experiment did. But we also found a significant effect of error directionality for gender which was not the case with Miltenburg's experiment. Again we believe this difference in evaluation of gender-related errors is because of the reasons explained above and are not influencing our conclusion that indeed English speakers react in a similar way as native Chinese speakers given error categories within image descriptions. In our data analysis, we run the same tests for each part of the survey. In every table pair (tables 5 and 6, tables 7 and 8, tables 10 and 11) the mean scores for each category and the results of the table in general, are always lower for the tables relating to the second part of the survey. That is a first indication that an image might influence participants perception when they judge correct descriptions with automatically generated ones. This information though, is not enough to draw conclusions regarding the second research question. Thus, we run some t-test to compare the means of the scores for the different error categories of the first part of the survey with the respective ones of the second part and these are shown by table 9. We found significant difference in only the clothing color-related errors of the two parts of the survey. In the other categories the difference was not significant. The fact that the two surveys had the same structure,i.e. the same order of questions, might have influenced the evaluations scores given in the second part of the survey, as some participants were likely to just recall their answers from the first part and put something around similar ranges in the questions of the second part. Since we found significant difference in at least one of the error types, clothing color-related errors, we can say that an image plays an important role on the perspective of the participants on the their description quality evaluation and needs to be studied further.

6.1 Explaining the results

It is now clearer, that people, either Chinese speakers or English speakers, attach different levels of importance to different aspects of an image description. The fact that we performed two different parts for the survey, one with pictures and one without, validated our hypothesis that the clothing color-related errors are perceivable at a glance. The prominent features of an image elicit strong responses and comparing results of the first part with the second part of the survey confirmed this theory that was not taken into consideration by Miltenburg et al. (2020). Further the color errors were blatant, i.e. black to red. One might considered a color error for example, describe the color as orange instead of red, not severe but since we made sure that color errors were blatant then the clothing color-related error scores were somewhat expected. Although the mean scores of gender-related errors for the first and second part of the survey were not significantly different, they were extremely low compared to the other error types. Their social relevance might elicit strong responses but also the fact that is 50-50 chance of getting it wrong, its either male or female. This was the only error type category were the acceptable answers are only two, thus is judged more harshly.

6.2 Limitations

As stated in the procedure section, the data collection did not go as planned and some adjustments were necessary to be made. Some people dropped out before finishing the survey and also many had hard time understanding the questions and what they were expected to do. After the first feedback received at the pilot study, the question was made clearer but again some people seem to find it quite hard to understand. The expectation from them was to evaluate with a low mark a description's quality whenever they thought that the error presented was a serious one. It could be seen as a drawback for the experiment as if you forget the instructions given then by human nature the first association you would make when evaluate how serious an error is, is by high marks. Even though at each question above the slider bar, it was stated that you drag the slider in order to evaluate the quality of the description and not how serious the error was, some

participants had a hard time. A huge downside for the experiment is that the survey is very repetitive and takes quite long. Because of the repetitiveness of the tasks, the Qualtrics platform was giving warnings that some answers show trends that are followed by bots. When checking those answers, it was not possible to tell if the answers were provided by human or by a bot, as they were all valid so were included in the data analysis. An interesting feature that the platform provides is that it has a duration section so we could detect which participant actually read and did the whole experiment. If a participant finished the whole experiment in 2 minutes, then it was clear that they did not read each question carefully and answer honestly. As the survey was provided through online platforms, the reactions of participants could not be recorded and also we could not exploit the duration feature at its fullest. This will further be analysed in the Future research section below. To wrap it up with the qualtrics platform there was one problem that was discovered after the data collection was completed. As can be seen by figure 5, the slider bar was preset to 0. When a participant wanted to mark a question with zero, then this was only possible if he/she touches and drags the slider to zero. Several responses were not recorder, because participants saw the slider at position zero and just moved on to the next question, and these gaps had to be manually filled. For future experiments, this problem could be resolved by specifying that you need to touch the slider at every answer in order for it to be recorded. By doing this the researcher could also check if the participants actually read and understand the questions. Relating to the experiment itself, there were some limitations as well. Our aim for the second research question was to test the effect a picture has in human judgement. The structure of the experiment might have limited this attempt as the order of questions in the two parts was the same. Once participants recognised that they could easily just ignore the picture in the second part of the survey and evaluate with similar scores as the ones entered in the respective questions of the first part of the survey. Adding to this after evaluating several questions there was a risk of a new pattern to emerge. The clear aim of the first part of the survey was to evaluate the severity of each error type based on the whole sentence provided and for the second part to evaluate based on the picture as

well. Due to the repetitiveness of the survey and how the questions were ordered, after a few evaluations, participants might skip reading the whole sentence or looking at the image once they have recognised what error type is presented to the question and evaluate solely based on their memory and past evaluations of that certain error type. In general one might say that is regular for an error type, e.g. gender error, to have similar range evaluations in every scenario but in order to conclude that we have to made sure that the participants actually looked at the picture and not skip it. Because the participants were recruited through social media, it was not possible to check whether they have actually looked at the picture or read the whole instruction and descriptions. Further, since the participants were all relatives or friends with the experimenter, the genuineness of their answers is questionable since we cannot be certain if their answers are truthful or encounter this survey as compulsory due to a relationship bound that they might have with the experimenter. Finally, the number of pictures used (7 in total) might not be considered sufficient to draw important conclusions on the effect a picture has in our perception of an error type's severity. More pictures would have exposed participants to different scenarios and would allow further research and stronger arguments to be made.

6.3 Future Research

Human Reactions. The experiment was performed through social media and was not possible to monitor on real time the participants reactions. The only thing possible was to have a discussion with them at the end but would be interesting to see what reaction they have when reading each question. Our intuition is that by monitoring the participant's reactions could be proven useful as to explain why we see such differences in error severity.

Error types. For the present study, we only considered four different types of errors. Future research should look into different kind of errors and see if there are again difference in the error severity once we increase the error categories. For example, gender-related errors had only two options, either male or female and would be interesting if we find a similar error type (with only two acceptable answers) and see if it yields similar strong responses.

Image Influence. For this experiment, we tested the difference in evaluations scores when having and not having a picture. To understand better the influence a picture makes in our evaluation process, we could test the difference of having a glance of a picture and having it for several minutes. When you have a glance of a picture, you only recall the most salient parts of it and evaluate accordingly. But when you stare to a picture for several minutes, you get to see every detail of it and judge more thoroughly.

7 Conclusion

In this study, we started our investigation with the problem of evaluation metrics and their poor correlation with human evaluations. Following an experiment on Chinese speakers carried by Miltenburg et al. (2020) we manage to overall replicate their results on English speakers proving that different kinds of errors elicit significantly different evaluation scores. These differences are unable to be captured by current evaluation metrics. Furthermore, we sought to investigate the influence a picture makes when evaluating the quality of descriptions. We presented the same two surveys, one with pictures and one without, but with the same question order and found out that pictures indeed influence our perspective. The results obtained suggest that salient features of a picture elicit strong responses while the presence of a picture allows for lower evaluation scores. More specifically, clothing color errors elicit stronger responses once a picture is present with participants evaluating with lower scores such errors. Moreover we found some evidence for differences within error categories when there is a picture present. This evidence pave the way for further research in a possible introduction of weighted quality metrics, i.e. evaluation metrics considering different levels of error severity by weighing different kinds of errors.

References

[1] Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *CoRR*, abs/1601.03896, 2016.

[2] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. What do we perceive in a glance of a real-world scene? *J. Vis.*, 7(1):10, January 2007.

[3] Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of NLG systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy, April 2006. Association for Computational Linguistics.

[4] Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for NLG systems. *CoRR*, abs/2008.12009, 2020.

[5] Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *CoRR*, abs/1703.09902, 2017.

[6] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. *CoRR*, abs/1612.07600, 2016.

[7] Marta Aracil Muñoz. A deep learning approach for automatically generating descriptions of images containing people. 2018.

[8] EHUD REITER and ROBERT DALE. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[9] Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin, Yibo Sun, and Ting Liu. Topic-to-essay generation with neural networks. pages 4078–4084, 07 2018.

[10] Lin Qiao, Jianhao Yan, Fandong Meng, Zhendong Yang, and Jie Zhou. A sentiment-controllable topic-to-essay generator with topic knowledge graph. *CoRR*, abs/2010.05511, 2020.

[11] Wei Wang, Hai-Tao Zheng, and Zibo Lin. Self-attention and retrieval enhanced neural networks for essay generation. pages 8199–8203, 05 2020.

[12] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

[13] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799, 2020.

[14] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision – ECCV 2010*, pages 15–29, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[15] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating simple image descriptions. In *CVPR 2011*, pages 1601–1608, 2011.

[16] Yezhou Yang, Ching Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 444–454, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[17] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[18] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. pages 529–545, 09 2014.

[19] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. 47(1):853–899, may 2013.

[20] Micah Hodosh and Julia Hockenmaier. Sentence-based image description with scalable, explicit models. In *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 294–300, 2013.

[21] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[22] Chen Sun, Chuang Gan, and Ram Nevatia. Automatic concept discovery from parallel text and visual corpora. *CoRR*, abs/1509.07225, 2015.

[23] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311, 05 2018.

[24] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[25] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[26] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China, 22–24 Jun 2014. PMLR.

[27] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemelauthor. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[28] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[29] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *CoRR*, abs/1212.4522, 2012.

[30] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311, 05 2018.

[31] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *CoRR*, abs/1406.5679, 2014.

[32] Semih Yagcioglu, Erkut Erdem, Aykut Erdem, and Ruket Cakici. A distributed representation based query expansion approach for image captioning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 106–111, Beijing, China, July 2015. Association for Computational Linguistics.

[33] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. *CoRR*, abs/1704.03899, 2017.

[34] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *CoRR*, abs/1612.00563, 2016.

[35] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M. Hospedales. Actor-critic sequence training for image captioning. *CoRR*, abs/1706.09601, 2017.

[36] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[37] Xingyi Song, Trevor Cohn, and Lucia Specia. Bleu deconstructed: Designing a better mt evaluation metric. *International Journal of Computational Linguistics and Applications*, 4(2):29–44, 2013.

[38] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[39] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[40] Yvette Graham and Timothy Baldwin. Testing for significance of increased correlation with human judgment. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 172–176, Doha, Qatar, October 2014. Association for Computational Linguistics.

[41] Natalie Schluter. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain, April 2017. Association for Computational Linguistics.

[42] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014.

[43] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. *CoRR*, abs/1607.08822, 2016.

[44] Naeha Sharif, Lyndon White, Mohammed Bennamoun, and Syed Afaq Ali Shah. Learning-based composite metrics for improved caption evaluation. In *Proceedings of ACL 2018, Student Research Workshop*, pages 14–20, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[45] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Optimization of image description metrics using policy gradient methods. *CoRR*, abs/1612.00370, 2016.

[46] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675, 2019.

[47] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[48] Michael Hanna and Ondřej Bojar. A fine-grained analysis of BERTScore. In *Proceedings of the Sixth Conference on Machine Translation*, pages 507–517, Online, November 2021. Association for Computational Linguistics.

[49] Ehud Reiter and Anja Belz. An Investigation into the Validity of Some Metrics for Automatically Evaluating Natural Language Generation Systems. *Computational Linguistics*, 35(4):529–558, 12 2009.

[50] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[51] Emiel van Miltenburg, Wei-Ting Lu, Emiel Krahmer, Albert Gatt, Guanyi Chen, Lin Li, and Kees van Deemter. Gradations of error severity in automatic image descriptions. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 398–411, Dublin, Ireland, December 2020. Association for Computational Linguistics.

[52] Emiel van Miltenburg, Miruna-Adriana Clinciu, Ondrej Dusek, Dimitra Gkatzia, Stephanie Inglis, Leo Leppänen, Saad Mahamood, Emma Manning, Stephanie Schoch, Craig Thomson, and Luou Wen. Underreporting of errors in nlg output, and what to do about it. In *INLG*, pages 140–153, 2021.

[53] SUN Mei and TIAN Zhao-xia. The cultural differences between english and chinese courtesy languages. *Journal of Literature and Art Studies*, 7(3):340–344, 2017.

[54] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg. Studying relationships between human gaze, description, and computer vision. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 739–746, 2013.

[55] Laurent Itti and Christof Koch. Computational modelling of visual attention, Mar 2001.

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

[57] Andy Field. *Discovering statistics using IBM SPSS statistics*. Sage Texts, fifth edition, November 2019.

8 Descriptions

Figures 6-12 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 Translation: A boy in a black shirt pitches at the baseball field

Gender Error: A girl in a black shirt pitches at the baseball field

Age Error: A man in a black shirt pitches at the baseball field

Clothing type error: A boy in a black coat pitches at the baseball field

Clothing color error: A boy in a pink shirt pitches at the baseball field

Figure 6: Image 320785 from MS COCO

Correct Translation: A man in a yellow shirt plays tennis on the tennis court

Gender Error: A woman in a yellow shirt plays tennis on the tennis court

Age Error: A boy in a yellow shirt plays tennis on the tennis court

Clothing type error: A man in a yellow coat plays tennis on the tennis court.

Clothing color error: A man in a purple shirt plays tennis on the tennis court.

Figure 7: Image 344149 from MS COCO

Correct Translation: A girl in a blue dress stands by the water park.

Gender Error: A boy in a blue dress stands by the water park

Age Error: A woman in a blue dress stands by the water park

Clothing type error: A girl in a blue suit stands by the water park

Clothing color error: A girl in an orange dress stands by the water park.

Figure 8: Image 372182 from MS COCO

Correct Translation: A man in a gray shirt stands on the street.

Gender Error: A woman in a gray shirt stands on the street.

Age Error: A boy in a gray shirt stands on the street.

Clothing type error: A man in a gray coat stands on the street.

Clothing color error: A man in a yellow shirt stands on the street.

Figure 9: Image 141759 from MS COCO

Correct Translation: A woman in a pink skirt throws a frisbee on the grass

Gender Error: A man in a pink skirt throws a frisbee on the grass.

Age Error: A girl in a pink skirt throws a frisbee on the grass.

Clothing type error: A woman in pink pants throws a frisbee on the grass.

Clothing color error: A woman in a blue skirt throws a frisbee on the grass.

Figure 10: Image 137767 from MS COCO

Correct Translation: A man in a black suit takes a selfie in the toilet

Gender Error: A woman in a black suit takes a selfie in the toilet

Age Error: A boy in a black suit takes a selfie in the toilet

Clothing type error: A man in a black swimsuit takes a selfie in the toilet

Clothing color error: A man in a white suit takes a selfie in the toilet

Figure 11: Image 218368 from MS COCO

Correct Translation: A woman in black shorts plays tennis on the tennis court.

Gender Error: A man in black shorts plays tennis on the tennis court.

Age Error: A girl in black shorts plays tennis on the tennis court.

Clothing type error: A woman in black trousers plays tennis on the tennis court.

Clothing color error: A woman in red shorts plays tennis on the tennis court.

Figure 12: Image 35948 from MS COCO