

# Generating Unambiguous, Natural and Diverse Referring Expressions

by

NIKOLAOS PANAGIARIS



Thesis submitted in partial fulfilment of the requirements  
of Edinburgh Napier University, for the award of  
***Doctor of Philosophy***

School of Computing  
Edinburgh Napier University

OCTOBER 2022

## *Author's declaration*

---

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Nikolaos Panagiaris, October 2022

# *Abstract*

---

Referring expression generation (REG) aims at generating natural language definite descriptions for objects within images called referring expressions (REs). Despite the substantial progress in recent years, REG models are still far from being perfect. Existing attempts focus exclusively on how accurately referring expressions describe an object. However, other essential natural language attributes such as diversity and naturalness are overlooked. Therefore, this thesis aims to develop REG systems that produce REs that are: (1) unambiguous: the generated sentences describe the object unambiguously; (2) natural: the REs should be less distinguishable from the human ones; (3) diverse: the REG model should be able to produce a set of REs for a given target object that are notably different.

A limitation of the language models that have been used in REG is that, they utilize a static global visual representation that is excessively compressed and lacks in granularity since all the visual information is fused into a single vector. Therefore, the first contribution of this thesis is a novel object attention mechanism that dynamically uses salient object features. To further demonstrate the advantages of attention in REG, a novel transformer model is proposed that exploits different levels of visual information.

Secondly, neural approaches that follow the encoder-decoder architecture are usually trained to maximize the likelihood of the generated word given the history of generated words. However, two shortcomings stem from this training scheme: (1) the exposure bias: the model is never exposed to its own error during training; (2) training-evaluation mismatch: during training a strictly word-level loss is used, while at test time the model is evaluated on sequence level metrics. Recently approaches that utilize reinforcement learning techniques have shown promising results in training neural systems directly on non-differentiable metrics for the task at hand. Thus, a second contribution that this thesis makes, is a novel optimization approach to REG based on the REINFORCE algorithm that normalizes the reward by averaging over multiple-samples. However, it was found that, while directly optimizing the evaluation metrics the models achieve higher scores, the generated text lacks diversity due to repeated n-grams. Thus, this thesis proposes the use of minimum risk training (MRT) as an alternative way of optimizing REG systems on sequence level.

Finally, to overcome the lack of diversity it is proposed to extend the investigation in generating sets of referring expressions. Specifically, the effect of different decoding strategies is investigated by comparing their performance along the entire quality-diversity space.

# TABLE OF CONTENTS

---

<b>AUTHOR’S DECLARATION</b>	<b>i</b>
<b>ABSTRACT</b>	<b>ii</b>
<b>TABLE OF CONTENTS</b>	<b>iii</b>
<b>LIST OF TABLES</b>	<b>vi</b>
<b>LIST OF FIGURES</b>	<b>viii</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Research questions and Contributions . . . . .	3
1.2 Publications . . . . .	6
1.3 Thesis Overview . . . . .	7
<b>2 LITERATURE REVIEW</b>	<b>8</b>
2.1 Traditional REG approaches . . . . .	8
2.1.1 Account for diversity in traditional approaches . . . . .	10
2.2 Neural REG approaches . . . . .	13
2.3 Inference for conditional language models . . . . .	16
2.4 Reinforcement Learning in NLG . . . . .	19
2.5 The evolution of attention mechanisms . . . . .	21
2.6 Conclusions . . . . .	25

---

<b>3</b>	<b>LANGUAGE MODELS</b>	<b>26</b>
3.1	Contributions . . . . .	27
3.1.1	LSTM-based language model . . . . .	28
3.1.2	Object attention language model . . . . .	29
3.1.3	Transformer . . . . .	31
3.1.4	Training objective . . . . .	34
3.2	Experimental design and results . . . . .	35
3.2.1	Implementation Details . . . . .	35
3.2.2	Evaluation . . . . .	36
3.2.3	Datasets . . . . .	37
3.2.4	Attention-based REG results . . . . .	38
3.2.5	Transformer-based REG results . . . . .	40
3.2.6	Comparison between the proposed language models . . . . .	42
3.3	Conclusions . . . . .	43
<b>4</b>	<b>SEQUENCE LEVEL TRAINING OBJECTIVES FOR REG</b>	<b>45</b>
4.1	Contributions . . . . .	46
4.2	Training REG with Reinforcement Learning . . . . .	47
4.2.1	Self-critical sequence training (SCST) . . . . .	49
4.2.2	REINFORCE with multiple-samples per data point: . . . . .	50
4.2.3	Reward Configuration . . . . .	51
4.3	Minimum Risk Training for Referring Expression Generation . . . . .	52
4.4	Combined objectives . . . . .	53
4.5	Experimental design and results . . . . .	54
4.5.1	Implementation Details . . . . .	54
4.5.2	Evaluation . . . . .	54
4.5.3	Evaluating different RL training configurations . . . . .	54
4.5.4	Results of the proposed RL objective . . . . .	56
4.5.5	Evaluating Minimum Risk Training for REG . . . . .	57
4.5.6	Comparison between objectives . . . . .	58

---

---

4.6	Conclusions . . . . .	60
<b>5</b>	<b>DECODING STRATEGIES</b>	<b>63</b>
5.1	Contributions . . . . .	64
5.1.1	Maximization-based decoding methods . . . . .	64
5.1.1.1	Greedy Decoding . . . . .	64
5.1.1.2	Beam Search . . . . .	65
5.1.1.3	Diverse Beam search . . . . .	65
5.1.2	Sampling-based decoding methods . . . . .	66
5.1.2.1	Sampling with temperature . . . . .	66
5.1.2.2	Top- <i>k</i> Sampling . . . . .	67
5.1.2.3	Nucleus Sampling . . . . .	67
5.2	Experimental design and results . . . . .	68
5.2.1	Implementation Details . . . . .	68
5.2.2	Evaluation . . . . .	68
5.2.3	Results for Random Sampling-based Decoding Methods . . . . .	69
5.2.4	Maximization-based Decoding Methods . . . . .	71
5.2.5	Human evaluation of sets of Referring Expressions . . . . .	73
5.3	Conclusions . . . . .	79
<b>6</b>	<b>CONCLUSIONS AND FUTURE DIRECTIONS</b>	<b>81</b>
6.1	Contributions and Findings . . . . .	81
6.2	Limitations and Future Work . . . . .	85
	<b>REFERENCES</b>	<b>87</b>
	<b>APPENDIX A HUMAN EVALUATION QUESTIONNAIRES</b>	<b>103</b>

# LIST OF TABLES

---

TABLES	Page
3.1 Comparison of different automatic metrics for the attention model (denoted as “LSTM-ATT”) and the standard LSTM model. The proposed attention model results in significantly higher CIDEr and $BLEU_1$ scores in both datasets. The p-values are the result of two-tailed t-tests using paired samples. . . . .	39
3.2 The impact of depth in the performance of the transformer model. Transformer 6 and 3 indicate that the decoder and the decoder consist of 6 and 3 layers respectively. The layer configuration follows the one proposed by Vaswani et al. [94]. “Ours” indicates that each layer of the encoder is connected with the respective layer of the decoder. . . . .	41
3.3 Human Evaluation results. Median scores for systems, mean and standard deviation in parentheses. . . . .	43
4.1 Performance of different reward functions for the LSTM model. When the language model is optimized with the CIDEr metric, a significant increase to all other evaluation metrics is observed. All models were decoded using greedy decoding. The performance of the seed model is also reported. The best overall values for each metric are emphasized with bold. . . . .	55

---

4.2	Results of different search strategies for reward computation and variance reduction for the LSTM model. “RS” stands for random sampling, while “BS” refers to beam search and “GD” for greedy decoding. “SCTS” refers to self-critical training. Shaping denotes that we used reward shaping. . . . .	55
4.3	Performance of the best attention (denoted as LSTM +ATT) and transformer model (denoted as Transformer ) trained with maximum likelihood estimation (denoted as MLE), self-critical sequence training (denoted as SCST) and the proposed RL objective (denoted as RL (OURS)). The models were greedily decoded. . . . .	57
4.4	System results for RefCOCO : CIDEr and BLEU scores; average sentence length (ASL); vocabulary size (Voc); mean-segmented bigram ratio (TTR); RL (ours) denotes the proposed RL objective; MRT denotes minimum risk training. . . . .	59
4.5	System results for RefCOCO+ : CIDEr and BLEU scores; average sentence length (ASL); vocabulary size (Voc); mean-segmented bigram ratio (TTR); RL (ours) denotes the proposed RL objective; MRT denotes minimum risk training. . . . .	59
5.1	The hyperparameter that controls the quality-diversity trade-off for each of the decoding strategies used in this work. . . . .	68
5.2	Hyperparameter configurations used in our human evaluation for each of the decoding strategies. . . . .	74



# LIST OF FIGURES

---

<b>FIGURES</b>	<b>Page</b>
1.1 Generated referring expressions by systems presented by (63, 115, 116) . . .	2
3.1 Overview of the proposed object attention language model. . . . .	30
3.2 Overview of the transformer architecture (94). The red arrows illustrate the original connectivity between the encoder and decoder, while the green arrows illustrate the proposed connectivity. . . . .	32
3.3 Human written referring expression for one target object (green box) for each of the test sets of RefCOCO and RefCOCO+. . . . .	38
3.4 Examples of objects and expressions drawn from RefCOCO dataset, for which the CIDEr scores of the attention model show an improvement over the standard LSTM. The target object is highlighted with a red box. . . . .	40
3.5 Examples of objects and expressions drawn from RefCOCO+ dataset, for which the CIDEr scores of the attention model show an improvement over the standard LSTM. The target object is highlighted with a red box. . . . .	40
3.6 Examples of objects and expressions drawn from both RefCOCO and RefCOCO+ datasets, for which the CIDEr score for the proposed transformer model show an improvement over the standard transformer. The target object is highlighted with a red box. . . . .	42

---

4.1	An illustration of the self-critical sequence training approach. In particular, a specific action $\hat{y}_2$ is sampled and the greedy action $\hat{y}_2^g$ is extracted. The difference of the rewards from sampling and greedy sequence is used to update the loss function. . . . .	49
4.2	Gradient variance of the proposed RL objective compared to the SCST for the proposed attention and transformer model. . . . .	56
4.3	Validation set CIDEr scores for different candidate set sizes for the MRT model. Best viewed in color. . . . .	58
4.4	Examples of objects and expressions drawn from both RefCOCO and RefCOCO+ datasets. The target object is highlighted with a red box. . . . .	60
5.1	An example image associated with the top three referring expressions decoded with standard beam search and those provided by humans annotators. The target object is highlighted with the green box. . . . .	64
5.2	Self-CIDEr and average CIDEr scores for random sampling with different temperature values for RefCOCO dataset. The language model used is the proposed transformer trained with cross-entropy (XE) and fine-tuned with the proposed RL objective. . . . .	70
5.3	Self-CIDEr and average CIDEr scores for random sampling with different temperature values for RefCOCO+ dataset. The language model used is the proposed transformer trained with cross-entropy (XE) and fine-tuned with the proposed RL objective. . . . .	71
5.4	Self-CIDEr and average CIDEr scores for top- $k$ and nucleus sampling for varying $k$ and $q$ values for RefCOCO dataset. The temperature was set to $T = 1$ . The language model used is the proposed transformer trained with cross-entropy. . . . .	72
5.5	Self-CIDEr and average CIDEr scores for top- $k$ and nucleus sampling for varying $k$ and $q$ values for RefCOCO+ dataset. The temperature was set to $T = 1$ . The language model used is the proposed transformer trained with cross-entropy. . . . .	73

---

---

5.6	Examples of objects and sets of expressions drawn from RefCOCO and RefCOCO+ datasets decoded with random sampling with varying temperature values. Human written expressions are also presented. . . . .	74
5.7	Self-CIDEr and average CIDEr scores for beam search and diverse beam search with varying temperature and diversity strength values for RefCOCO dataset. . . . .	75
5.8	Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO testA. . . . .	76
5.9	Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO testB. . . . .	76
5.10	Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO+ testA. . . . .	77
5.11	Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO+ testB. . . . .	78
5.12	Examples of objects and sets of expressions drawn from RefCOCO and RefCOCO+ datasets. The expressions were decoded with beam search and diverse beam search. . . . .	79
A.1	Amazon mechanical turk example question for evaluating the quality of a single referring expression. . . . .	104
A.2	Amazon mechanical turk example question for evaluating the quality and diversity of a set of referring expressions. . . . .	105

# Introduction

---

Referring Expression Generation (REG) aims at generating natural language descriptions for objects within scenes called referring expressions (REs) (48). The recently released datasets RefCOCO, RefCOCO+ and RefCOCOG (67, 116) which contain natural images of cluttered scenes impose new challenges to the task. Referring to objects in open domain images requires in depth understanding of the global concepts of the image, as well as their attributes and relationships. Deep learning approaches have yielded promising results on this task (8, 115, 116, 119). Such approaches derive their inspiration from the recently introduced encoder-decoder paradigm (72) originally proposed for machine translation (12, 87) and since have been widely used in various NLG sub-fields such as storytelling (19, 36), summarization (25, 91), dialogue systems (55, 103), and image captioning (102, 108). This architectural scheme utilizes a deep convolutional neural network (CNN) (49) to extract a vector representation of an image or image region, and a variation of recurrent neural networks (RNNs) (42), e.g. a Long Short-Term Memory (LSTM) network (34) to generate the output.

Despite the substantial progress in recent years REG models are still far from being perfect. Existing neural REG attempts focus mostly on the generation of unambiguous referring expressions. However, other essential natural language attributes such as *diversity* and *naturalness* have received less attention. Diversity is important for a number reasons. First, an image contains multiple concepts at various levels of detail, and thus a RE describes a set of attributes that are interesting to the human speaker that uttered



**Figure 1.1:** Generated referring expressions by systems presented by (63, 115, 116)

the expression. It has been shown that the content of a RE is speaker dependent (97). In other words for the same referential environment (e.g. image), different speakers will often utter diverse expressions, a property that is reflected by the naturally existing human text. However, as it shown in Figure 1.1 generated expressions are repetitive. Furthermore, each of the REG datasets used in this study, namely RefCOCO and RefCOCO+, average 3 REs per object. Hence, from a machine learning standpoint, it is reasonable not only to evaluate the modes of the learned conditional distribution that reflect the accuracy, but also its variance which reflects the diversity of the generated output (104).

Furthermore, as robots become increasingly pervasive in modern societies, it becomes increasingly important to endow them capabilities that allow fluent and natural human-robot communication. Attracted by the naturalness of human communication, robots were equipped with natural language capabilities especially those designed to operate in domains that require cooperation or communication with human users. For a single-purpose robot such as a vacuum cleaner, this interaction can be simplified even to the press-of-a-button command. However, controlling robots that perform complex tasks requires advanced communication capabilities, including the ability to refer to objects or events in dynamic environments. For instance, users of housekeeper robot that can fetch and deliver objects need to be able to refer to an arbitrary target object and delivery point. More importantly, in the near future, new technological advances might take into account all ongoing visual aspects of the environment in real time. Such technological improvement will permit multipurpose human assistive robots to guide

visually impaired and elderly people to some predefined destination avoiding obstacles and traffic. However, such communication scenario is highly dynamic and the robot needs to refer to objects and events with high precision and be able to produce a set of diverse responses in case the user is unable to understand a particular description.

Therefore, in this thesis an alternative approach is explored as to what a “good” referring expression is. The objective of the systems developed in this thesis is to produce referring expressions that are: (1) *unambiguous*: the generated expressions should describe the object univocally; (2) *natural*: the referring expressions should be less distinguishable from the human ones; (3) *diverse*: the REG model should be able to produce a set of referring expressions for a given target object that are notably different.

## 1.1 Research questions and Contributions

This section states the research questions explored in this thesis and describes the approaches that were developed to address those questions.

- **RQ1:** *To what extent do language models affect the ambiguity and naturalness of referring expressions?*

First, it is proposed to incorporate a novel object attention mechanism to the standard RNN network that has been used so far in REG. Under the standard RNN framework, the generation of the next word is conditioned on the previously generated words. While this may suffice when the visual stimuli is relatively simple, for complex cluttered scenes a more fine-grained visual representation is required in order to generate high quality output. The attention mechanism bridges this gap by learning to focus on regions that are salient. The key novelty of this model is that, instead of letting the language model to hallucinate over the attributes that sound plausible the attention mechanism enables the language model to be exposed to multiple salient regions of the object during generation. It is shown that the inclusion of the proposed attention mechanism has significant benefits for REG.

To further demonstrate the benefits of attention in neural REG, a transformer-based (94) model is proposed. Transformers have revolutionized NLG fields such as machine translation, where the machine generated translations surpass the performance of those produced by human experts (94). However, there are limited attempts to incorporate the transformer models in vision & language tasks. To bridge this gap, this thesis investigates the effectiveness of the original architecture and it proposes a novel layer configuration in order to provide the network with a global “context” signal by connecting each layer of the encoder with the respective layer of the decoder. It is shown that the proposed transformer model is highly effective. Specifically, significant improvements are reported, both quantitative and qualitative, over baseline methods and the results compare favorably to the state-of-the-art results not only in automatic metrics but also in human evaluation.

- **RQ2** *To what extent do training objectives affect the ambiguity, naturalness and diversity of the referring expressions?*

The encoder-decoder models are trained mostly to maximize the likelihood of the generated word given the history of generated words that far. This approach has been coined in literature as “Teacher-Forcing” (4). A limitation that stems from this approach is that the model is never exposed to its own predictions during training, while during generation the model uses its own predictions to generate the next word. Furthermore, there is a loss-evaluation metric mismatch coined as *exposure bias* (79). During training the model utilizes a word-level loss, while during generation its goal is to generate an expression that improves sequence-level metrics. Recently, a completely novel point of view has emerged in addressing these two problems, that is utilizing reinforcement learning techniques (88). However, training with RL is a non-trivial task due to a number of limitations: (1) high variance of the gradient (81); and (2) reward configuration (79).

A limitation that stems from RL based approaches is that usually they utilize one sample per data point. This thesis argues that one sample might be insufficiently

expressive for an observation. As a result, samples that poorly describe the observation will be heavily penalized, pushing the model to cover only high-probability zones. To minimize this effect, this thesis propose an effective way to calculate the baseline of the REINFORCE algorithm. Specifically, the proposed method normalizes the reward by averaging over multiple-samples per observation. The underlying idea of this method is that drawing multiple diverse samples allows the construction of a robust baseline due to the diversity of the samples that are considered. In other words, averaging over multiple samples lifts the burden of having each sample to explain the observation well.

Furthermore, it was found that while directly optimizing the evaluation metrics one can achieve higher scores, the generated text lacks diversity due to repeated n-grams (105). The analysis shows that RL trained models are strongly biased towards frequent REs leading to smaller vocabulary and deficiencies in the generated word distribution. To address these issues the use of minimum risk training (MRT) (84) is proposed as an alternative way of optimizing REG systems on sequence level. Minimum risk training aims at minimizing the expected loss over training data by taking automatic evaluation metrics into consideration. The MRT objective has the following advantages over MLE. First, it can directly optimize sequence level objectives that are not necessarily differentiable. Second, while MLE maximizes the likelihood of the training data, MRT introduces a notion of ranking amongst candidate sequences by discriminating between sequences. Thus, by minimizing the risk, we expect to find a distribution that approximates well the ground-truth distribution. Furthermore, the MRT objective is similar to the REINFORCE algorithm in a sense that both maximize an expected reward or cost. However, there are two fundamental advantages of the MRT over RL: (1) the REINFORCE algorithm typically utilizes one sample in order to approximate the expectation, whereas the MRT objective considers multiple sequences making it sample and data sufficient; and (2) the MRT objective intuitively estimates the expected risk over a set of candidate sequences, whereas the REINFORCE algorithm



typically relies on the baseline reward to determine effectively the sign of the gradient. Finally, a detailed analysis shows that when a REG model is trained with the proposed approach, it uses a larger vocabulary, produces longer referring expressions and generates more uni-grams and bi-grams.

- **RQ3** *To what extent do decoding methods affect the ambiguity, naturalness and diversity of the referring expressions?*

To overcome the lack of diversity it is proposed to extend the investigation in generating sets of referring expressions. Specifically, the effect of different decoding strategies is investigated by comparing their performance along the entire quality-diversity space. The importance that NLG systems place on these two criteria, is application dependent. For example, the goal of an open domain dialogue generation system is to be able to converse for a variety of topics and thus places more weight in the diversity of the output (51). However, in REG the most important attribute of the output is to successfully identify the target object. Thus, generating a set of expressions is useful only if it does not come on the expense of the quality. Therefore, the first large-scale human evaluation is presented in order to measure how the hyperparameters of each decoding algorithm, affect the diversity and the quality of sets of referring expressions.

## 1.2 Publications

The work presented in this thesis has been originally published in:

- Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. Generating unambiguous and diverse referring expressions. *Computer Speech Language*
- Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. Improving the naturalness and diversity of referring expression generation models using minimum risk training. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 41–51, Dublin, Ireland, 2020. Association for Computational Linguist.

## 1.3 Thesis Overview

The remainder of the thesis is organised as follows:

- **Chapter 2** discusses the background; First, a review of traditional approaches in REG is given. Traditional approaches that accounts for diversity are also discussed. Second, a review of Neural REG approaches is given. Third, an overview of different ways to implement inference in conditional language models is given. Fourth, a survey of works that leveraging methods from reinforcement learning is presented. Fifth, the evolution of attention mechanisms is presented. Finally, the chapter concludes with a critical analysis of current practices
- **Chapter 3** develops and compares two language models for REG. The first model incorporates a novel object attention mechanism to the standard RNN network. The second model is transformer based architecture with a novel layer configuration in that connects each layer of the encoder with the respective layer of the decoder. The benefits and limitations of each approach are elaborated.
- **Chapter 4** develops and compares two sequence level objectives for REG. The first approach is a novel optimization approach to REG based on the REINFORCE algorithm. The second approach, is a novel strategy for training REG models using minimum risk training. The benefits and limitations of each approach are elaborated.
- **Chapter 5** investigates the generation of sets of referring expressions and the effect of different decoding strategies by comparing their performance along the entire quality-diversity space.
- **Chapter 6** summarises the main findings and contributions of this thesis and suggests possible avenues for future work.

## *Literature Review*

---

This chapter presents an overview of related work to this thesis. Specifically, the chapter begins by presenting traditional content selection approaches in Section 2.1. Section 2.1.1 describes traditional approaches that accounts for diversity. Next, Section 2.2 describes Neural REG approaches that is the focus of this thesis. Section 2.3 provides an overview of the different ways of implementing decoding on top of neural network-based generation models. Section 2.4 presents a survey of works that leveraging methods from reinforcement learning. Section 2.5 presents the evolution of attention mechanisms. Finally, Section 2.6 concludes the chapter by critically discussing the different approaches.

### **2.1 Traditional REG approaches**

Traditionally, REG systems have been seen as a multi-step process that includes a number of choices in order to transform the input to a natural language description. The first choice is which form a referring expression will assume, i.e. whether the target object will be referred to with a proper name, a definite description or a pronoun. If the chosen form is a description, the second step is the determination of the content, that is the selection of properties that distinguish the target object from potential distractors (i.e. objects similar to the target) in a given context. The last step is the linguistic realisation of all the properties to a fully-fledged description. The large body of existing work in REG, focuses on the determination of content for definite descriptions (48).

Content selection algorithms search for a combination of properties that distinguishes the target object univocally. The termination criterion of the search depends on the modeler's interpretation of what constitutes a "good" referring expression. A large body of literature defines that a "good" referring expression is that which does not violate the Maxim of Quantity (15). In other words, a referring expression should convey just *enough* information to unambiguously identify the referent but no more. What constitutes "enough information" has led to a number of algorithmic definitions. Since such approaches are not the scope of this thesis the most studied algorithms are described.

First, the full brevity algorithm (14) exhaustively searches the space of possible properties of the referent in order to produce the smallest set that unambiguously identifies the referent. Specifically, the brevity algorithm first checks whether a target object can be identified by using only one attribute and if not, the algorithm searches for all possible combinations of two attributes and checks if any combination facilitates the identification of the target object and so on. However, (16) showed that finding the shortest distinguishing description is a NP-Hard problem. Thus, a greedy heuristic approach was proposed by Dale [14], which incrementally chooses the properties that rule out the most distractors in the domain, and thus minimising the possibility of including ambiguous information about the target object.

In order to address the computational complexity of finding the shortest possible distinguishing description the incremental algorithm (IA) was introduced (16). Like greedy search, the incremental algorithm selects properties one by one until the distinguishing description is formed. The main difference between the two approaches, is that the incremental algorithm instead of checking through the complete list of the truthful properties for the target object in each iteration and identify which one excludes the most distractors, it chooses a property that rules at least one potential distractor. In other words, the order in which the properties are added to the description is not based on their discriminatory power but is based on a predefined preference order which dictates in which order these properties need to be considered. The preference order drives its inspiration from psycholinguistics studies (e.g. (75)) that support the notion

---

of prominence of particular attributes compared to others, and are likely to be included in the descriptions.

### **2.1.1 Account for diversity in traditional approaches**

This section describes early work in traditional approaches that strives for human-likeness and diversity. Specifically, with the introduction of the data oriented methods an observation that was made is that each human speaker or writer prefers distinct properties, syntax and lexical units while building referring expressions. The existence of corpora allowed modelers to use corpus frequencies to find an algorithm that outputs expressions that mimic those produced by humans. Such approaches are either based on hand crafted rules or based on machine learning (ML) algorithms. Typically learning algorithms use as features, for example, the number of distractors, the number of objects having the same type, and so on. However, those features do not convey speaker-related information. To address this issue, the use of the speaker's identity (or speaker demographic features) as additional features that account for diversity was proposed.

[Bohnet \[5\]](#) presents a variation full brevity that uses a Nearest-Neighbor (13) learning technique to build an individual referring expression model for each speaker. They achieve that by selecting expressions that are similar to expressions produced previously by the same speaker. [6](#) proposes a variation of the incremental algorithm that uses corpus frequencies for each speaker to create individual preference lists, i.e. unlike the classical approach that creates domain specific preference lists, their approach orders the attributes by how frequently were used by each speaker. They showed that this ordering strategy improves the accuracy (i.e. the similarity at string level with the ground truth reference) of the selected attributes by 0.02 compared to the use of a global attribute list that is ordered based on the frequency of the complete training set. In terms of variation in lexicalization choices, different models of vocabulary and syntactic expressions are created for different speakers. Information concerning frequent choices made by speakers in the use of determiners, and syntactic patterns that are mostly preferred are considered.

A similar approach has been followed by [Di Fabbri, Stent and Bangalore \[18\]](#) in which extensions of full brevity and incremental algorithm are proposed. However, in the case of Full brevity algorithm this approach uses the attribute set that is used most often and most recently by each speaker. The extension of incremental algorithm is similar to that proposed by [Bohnet \[7\]](#) since it uses a speaker-specific ordered attribute list. They showed that incorporating speaker traits matters greatly in terms of accuracy. In order to capture the diversity in the choice of words and structures different samples that exist in the corpus are considered as templates. A problem that stems from this approach is that, if a set of attributes has never been used in the corpus, the system is unable to find a suitable template.

[Viethen and Dale \[98\]](#) presents a decision tree approach for the content selection of referring expressions. In particular, the C4.5 decision tree learning algorithm (77) was used. The decision tree that was learned (i.e the content pattern that can be predicted by that tree) can be expressed as the following rule: “ If there are a few distractors that share the same size as the entity in question, choose the content pattern R else use the pattern D”. The pattern D contains two attributes, the type and color, while R contains the size. However, those two rules proved to result in significant low accuracy, which is defined as the number of instances predicted correctly divided by the total number of instances in the test or training set (98). When authors introduced as a feature the identity of the participants, the resulting speaker-sensitive decision trees become much more complex but resulted in much higher accuracy scores. In fact, it was shown that even excluding all the scenes features and using speakers identity as the only feature, the accuracy of the resulting decision tree can be roughly compared with those learned with only the scene features. Tantalizingly, authors argued that future research should explore: (1) automatic clustering techniques for grouping people that use the same pattern in reference production, aiming to construct separate algorithms for each profile and (2) whether the use of non-linguistic characteristics of speakers (e.g. demographic features or cultural background), can account for some of the between participant variation in reference behavior.

The former was explored by (99) where an extension of a graph based algorithm was

---

introduced. In a nutshell, what differentiates this algorithm from the classical approach is that instead of searching for the cheapest distinguishing subgraph it searches for the longest of all the cheapest subgraphs. Furthermore, instead of using the speaker's identity as an input parameter to the system, authors proposed to automatically find clusters of speakers with the same referential behavior. In order to achieve that, they used K-means clustering which resulted into two speaker profiles. Specifically, the first profile contains descriptions from speakers that produce relatively long descriptions, while the second is consisted of speakers that produce short description. Furthermore, they trained the search algorithms with separate cost functions and preference orders for the different clusters. This setting resulted in significantly better results compared to a setting that does taking into consideration speaker groups characteristics. However, it would have been interesting if it has been explored whether grouping speakers outperforms the use of speaker identity features. This work was further extended by [Liu, Wang and Yang \[60\]](#) in order the model to take learned attributes for objects as an extra input to the LSTM model, so that the generated expression bears high similarity in terms of attributes with those contained in the training samples. To model differences between objects, the MMI objective (67) was used.

More recently, classifiers have been used for content selection to decide whether an attribute should be added in the referring expression. [Ferreira and Paraboni \[20\]](#) casts the problem of REG as a classification problem. Specifically, individual binary classifiers are used to decide whether each referential attribute (atomic or relational) of a given target object should be selected to appear in the final description by combining the output of each classifier. In order to take the issue of human diversity into account, authors admitted two kind of features: (1) enriched personal information about the speakers; they included not only speaker's identity but also the gender and age bracket of each speaker as suggested by [Viethen and Dale \[98\]](#); and (2) speaker's referential behavior features (i.e. lists of attribute frequencies). Once again, they found that classifiers that include speakers personal information outperforms those that don't in terms of accuracy.

[Hervas, Francisco and Gervás \[32\]](#) presents an analysis for TUNA corpus (17, 22) as

a case-based reasoning in order to demonstrate that the performance of a lexicalization algorithm that emulates the human-generated referring expressions can be improved greatly if takes into consideration the particular choices made by each speaker. In particular, a lexicalization algorithm was tested following two different approaches: one that considered the lexical choices of the person who had contributed referring expression in the corpus, and another that did not. The latter approach followed a general corpus based modeling for lexicalization preferences while the former considered only lexicalization choices of different speaker in the corpus. It was found that speakers tend to be consistent with their own set of choices. [Hervás et al. \[33\]](#) extended this idea by arguing that sets of choices can be shared by different groups of people.

## 2.2 Neural REG approaches

The Section 2.1 described work in computational models of referring expressions for objects in artificially generated scenes. Yet how people refer to objects in real-world cluttered scenes is a relatively unexplored area. Neural REG approaches ([64](#), [67](#), [115](#), [116](#), [119](#)), have seen a surge of interest due to the availability of larger and more complex REG datasets such as RefCOCO (+) ([116](#)) and RefCOCOg ([67](#)). As mentioned earlier, such approaches follow the encoder-decoder paradigm. The underlying idea of the encoder-decoder is the following: a convolutional neural network processes the image region in order to extract a vector representation that is used to initialize the decoder (e.g. a recurrent neural network). Given the previous generated words, the next word in the sentence is predicted sequentially.

[Mao et al. \[67\]](#) were the first to apply the encoder-decoder architecture in REG. In particular, they use a convolutional neural network to extract visual features and an LSTM network ([35](#)) to generate the expressions. Three kind of visuals features were used: (1) features from the bounding box around the target object; (2) features from the whole image to serve as context; and (3) the relative location and the size of the region were target object is placed within the image. The relative location and size of the region were encoded using a 5 dimensional vector as follows:  $[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{S_{bbox}}{S_{image}}]$ , where



$(x_{tl}, y_{tl})$  and  $(x_{br}, y_{br})$  are the coordinates of the top left and bottom right corners of the object bounding box,  $H$  and  $W$  are height and width of the image, and  $S_{bbox}$  and  $S_{image}$  are the sizes of the bounding box and image respectively. Their final representation was the concatenation of region, image, and location/size features, which resulted in a 2005-dimensional vector. Instead of training the with the maximum likelihood objective the proposed the used of the maximum mutual information (MMI) objective in order to reduce the ambiguity of the output. The underlying idea of this objective is to capture whether a listener would deference a referring expression. They enforced this intuition by penalizing the model if a generated expression is likely to be generated for other objects within the image. To further reduce the ambiguity of the produced expression they introduced a comprehension model that can re-ranked them in the post-process. The referring expressions were decoded with beam search with a width of three. Furthermore, in order to evaluate their approaches they conducted human evaluation. Specifically, for their best performing model they reported that the produced referring expressions have task success of 65.32% in RefCOCOg dataset (67).

Yu et al. [117], extended the visual representation proposed by Mao et al. [67] by hypothesizing that a general feature over the entire image may not be sufficient to capture visual comparisons between the target object and other objects of the same object category within the image. In order to address this issue, they proposed two types of features for visual comparison. The first type is concerned of capturing similarities and differences in appearance between the target object and its distractors. They experimented with selecting different subsets of comparison objects and different strategies for computing an aggregate vector to represent the visual difference between the target object and the surrounding objects. The second type of comparison features represents the relative location and size differences between the target object and neighboring objects of the same object category (unlike (67) that calculated the size and location of the region). They chose up to five comparison objects that are close and of the same category as target object. Their final representation is the concatenation of the above features. Furthermore, they proposed an encoder-decoder method that ties the language generation process together for all depicted objects that share the same type as

the target object within the image in order to reduce ambiguity. In order to evaluate their approach they performed human evaluation. Specifically, they reported that the 73.77% of the produced referring expressions for RefCOCO successfully identify the target object, while for RefCOCO+ 51% of the produced referring expressions successfully identified the target object.

Yu et al. [114] proposes a unified model that jointly learns both the speaker and an embedding-based listener model. For the speaker they followed the classical encoder-decoder paradigm. In particular, they proposed a new objective function that generalizes the idea underlying the MMI proposed by Mao et al. [67]. In particular, the constraint that encourages the generated expression to describe the target object unambiguously was the incorporation of two triplet hinge losses composed of a positive match and two negative matches. In other words, given a positive match they sampled the contrastive pair which is an expression that describes another object within the image. Additionally, they added a discriminative reward-based reinforcer to guide the sampling of more discriminative expressions and further increasing the discriminatory power of the produced REs. Rather than working independently, they allowed the speaker, listener, and reinforcer to interact with each other. In order to evaluate the effectiveness of the produced referring expressions they performed human evaluation. Specifically, for their best performing model they reported that the 77.52% of the produced referring expressions for RefCOCO images were successful in identifying the target object correctly, while they reported that the 58.58% of referring expressions successfully identified the target objects in RefCOCO+ images.

Luo and Shakhnarovich [63] departed from the encoder-decoder paradigm, and drew inspiration from the generator-discriminator structure in Generative Adversarial Networks (GANs) (24). In particular, the generator produces the expressions, while the discriminator plays the role of the comprehension model since it tries to identify if the expression can be dereferenced. Instead of the adversarial relationship between the two modules, the authors chose a collaborative one. They assigned to the discriminator the role of a "guide" in order to improve the output of their system. Specifically, they utilized the discriminator in two ways. The first is the generate-and-rerank method which

uses comprehension on the fly. In other words, candidate expressions are generated by the generator and then are passed to the comprehension model which picks the expressions with highest generation comprehension score. The second one, it is training by proxy. The generation and comprehension model are connected and the generation model is optimized to lower the discriminative comprehension loss (in addition to the cross-entropy loss). In order to evaluate the effectiveness of the produced referring expressions they performed human evaluation. Specifically, they reported that 70.5% of the produced referring expressions for RefCOCO were successful, while the 45% of the produced referring expressions for RefCOCO+ images successfully unidentified the target object.

### **2.3 Inference for conditional language models**

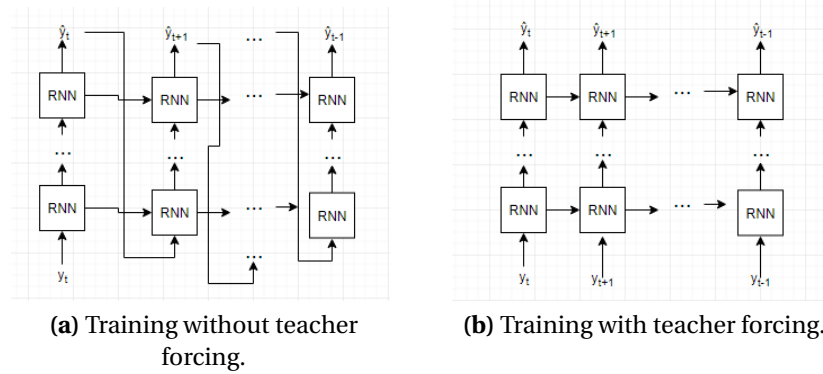
Despite recent efforts in modeling context and learning, decoding has received little attention, with the notable exception of, for example, (119). During inference all proposed methods in REG utilize a standard decoding algorithm, e.g. greedy search or beam search. Specifically, words that maximize the likelihood are drawn sequentially. However, what is the best decoding strategy for NLG models still remains an open challenge. Although the maximization of the likelihood as training objective produces high quality models, the maximization-based decoding algorithms produce text that is repetitive (37, 40, 100). A number of diversity promoting variants of beam search have been proposed for different NLG tasks. Specifically, the noisy parallel approximate decoding was proposed by Cho [10] for machine translation. Random noise is added to the hidden state of the decoder at each generation step. Specifically, noisy parallel approximate decoding is a meta-algorithm that runs in parallel multiple noisy versions of a traditional decoding algorithm, such as beam search. The final sequence candidate is the one with the highest score. Authors compared their proposed approach against the greedy search, beam search as well as stochastic sampling in an attention-based neural machine translation model trained on 12m sentence pairs, available from WMT'15 corpus. As their primary evaluation metric they used the negative conditional log prob-

ability (NLL) of a decoded sequence and as secondary the BLEU metric. They reported a reduction of 0.8 in NLL against greedy decoding and a reduction of 0.2 against beam search. Diverse beam search (100) was proposed for image captioning as it promotes diversity by penalizing new hypotheses that share same tokens with previously generated hypotheses. In order to achieve that, they partitioned the candidates into groups and they augmented the log probabilities of each word with a dissimilarity term. Empirically they found that the best performing dissimilarity factor is hamming diversity which penalises the word selection based on the frequency each word appears in previous groups. They tested their model on the MSCOCO dataset (59). In order to evaluate their approach they used the number of distinct n-grams where they showed an increase in the production of unique 4-grams against traditional beam of 300%. For machine translation and open dialog generation, top-g capping beam search was proposed by Li and Jurafsky [52], where only the top-g hypotheses from the same ancestor hypothesis are kept. They trained their models on the WMT German/English and French/English tasks. Furthermore, they evaluated the diversity of their systems by calculating the average number of distinct unigrams and bigrams. They found that their proposed diverse decoding produces 0.40% more distinct unigrams compared to beam search and 1.44% more unique bigrams for both for English-German translations and English-French translations. The iterative beam search, that was originally proposed for dialog generation, runs multiple iterations of beam search while excludes any previously explored space. To evaluate the diversity of the output of their proposed model they compute distinct n-grams over the set of the produced responses. They found that the output of their proposed method performs similarly with the traditional beam search. Finally, all the aforementioned share the same core idea of promoting diversity. Specifically, as shown such approaches include an additional factor in order to promote the expansion of diverse candidate expressions.

A strand of research investigates the augmentation of beam's search objective by training an additional network that provides a supplementary score to the likelihood. Specifically, Li, Monroe and Jurafsky [53] train an additional neural network to predict a reward for each partial hypothesis. Their model can be seen as a variation of the

actor-critic model that uses the language model as the actor and the trained model that estimates the future values of the candidate sequences plays the role of the critic. They used the OpenSubtitles (OSDb) dataset (93) to train their models. They reported an increase in the production of unigrams and bigrams compared to beam search. Similarly, Zariw&ouml; and Schlangen [119] treats the decoder as small actor networks that is trained to manipulate the hidden state of the underlying REG system. Furthermore, they explored a number of beam search variants proposed for machine translation that control the length of the produced expressions. They trained their models on RefCOCO and RefCOCO+ (116). They used three metrics to evaluate their models: (1) BLEU for unigrams, CIDEr, and length ratio. They reported that their trainable decoder improves over greedy decoding on almost test sets, models and measures. Finally, a potential advantage of approaches that train an additional module to handle the decoding of the sequences is that any objective or communication goal can be directly optimized by the additional module.

Another approach to the decoding step, is to sample from the model's learned distribution. Such approaches instead of framing the decoding problem as search problem they formulate it as a sampling problem. Specifically, under this scheme, at each time step, sampling-based decoding algorithms sample the next word by drawing a word from the conditional language model. While text generated by this method shows significant diversity, it can easily become incoherent because words from the model's less robust confidence areas can be drawn (36). To the best of our knowledge, three different ways have been proposed to address this issue: (1) the use of temperature to reduce the entropy of the distribution leading to a more skewed distribution towards the high confidence zones; (2) top- $k$  sampling (19), where a fixed number of  $k$  tokens is kept and the next word is sampled from this truncated vocabulary; (3) nucleus sampling (37), that keeps those tokens whose cumulative probability exceeds a pre-defined threshold. Such approaches alleviate the problem of drawing words from the model's less robust confidence areas by truncating the learned distribution in different ways.



## 2.4 Reinforcement Learning in NLG

The models presented in Section 2.2 are trained using a ground-truth sequence via a mechanism known as “Teacher-Forcing” (4), where the teacher is the ground-truth sequence. A limitation that stems from this approach is that the model is never exposed to its own predictions during training, while during generation the model uses its own predictions to generate the next word. Furthermore, there is a loss-evaluation metric mismatch coined as *exposure bias* (79). During training the model utilizes a word-level loss, while during generation its goal is to generate an expression that improves sequence-level metrics. Recently, utilizing methods from reinforcement learning (RL) has emerged as a solution for the two problems. This section aims to summarize such research in NLG relevant to this work.

The concept of improving the generation by exposing the model to its own predictions during training was first proposed by He et al. [29]. Specifically, they argued that the structured prediction problems can be cast as reinforcement learning problems. The underlying idea is to use the model’s predictions during training in order to produce sequences of actions (i.e. words). Then, a greedy search is used to determine the optimal action at each time step, and the policy is trained to predict that action. Ross, Gordon and Bagnell [82] proposes DAGGER an imitation learning approach, where actions given by an expert are required for each predicted word. A limitation that stems from this approach is that having actions given by an expert might not be feasible due to the large action space. Thus, Venkatraman, Hebert and Bagnell [96] proposes the “DAD”

model where the target action for each step is given by the optimal policy. A limitation that stems from this approach is that if the generated output is shorter in length than the ground truth sequence, the model will repeat previously generated actions.

The REINFORCE algorithm was proposed as solution to the two aforementioned problems. One of the earliest adoptions of REINFORCE algorithm was proposed by [Ranzato et al. \[79\]](#). In this method, first the language model is trained with the cross-entropy loss for  $N$  epochs utilizing the ground truth sequences. They argued that pretraining guides the model to focus on promising regions of the search and thus it results in a better policy. Then, they proposed “MIXER” an annealing schedule algorithm in order to gradually switched from the cross-entropy loss to REINFORCE loss to train the model. Later work that utilizes the REINFORCE algorithm is mostly focused on the calculation of the reward. Specifically, [Ren et al. \[80\]](#) compute a visual-semantic based similarity score that is used as reward. [Liu et al. \[62\]](#) proposes to use as a reward a linear combination of the SPICE (2) and CIDEr (95) metrics, called SPIDEr.

A major limitation that stems from the REINFORCE algorithm is that, the expected gradient exhibits high variance and without careful normalization is often unstable (81). An extension to the REINFORCE algorithm includes the reduction of the variance by subtracting a quantity from the learning signal called a baseline (83, 118). However, how to best calculate the baseline is not trivial and thus it led to a thread of research that investigates different ways to calculate the baseline. As one for the early attempts to effectively calculate the reward is the use of Actor-Critic models. Specifically, [He et al. \[27\]](#) utilizes a semantic matching and a context-coverage module to estimate the value function. However, in order the value network to be trained, a fully trained language model is required. Then, they utilize the trained language model and the trained value network during inference. Therefore, the value network is not used during the training of the language model. Similarly, [Li, Monroe and Jurafsky \[54\]](#) trains a value function approximator which estimates the future outcome of taking an action in the present and it incorporates it at each decoding step. [Li, Bing and Lam \[56\]](#) proposes an Actor-Critic based model where a binary classifier plays the role of the Critic. The role of the Critic in this model is to distinguish the generated output from that proposed by human writers.

During the training of the Actor-Critic model, the score from the binary classifier is used as a surrogate for the value function.

A limitation that stems from the Actor-Critic methods is that there is a need to estimate both action-dependent and action-independent rewards functions which usually involves the training of at least one more model. To overcome these problems, [Rennie et al. \[81\]](#) proposes a new approach called self-critical sequence training (SCST). In this model, rather than estimating the reward signal, or how the reward signal should be normalized, it utilizes the output of the model obtained by a greedy-search (the output at the time of inference). As a result, samples that outperform the current system are given positive scores, while inferior samples are penalised. In other words, SCST elegantly avoids to estimate the reward signal (as Actor-Critic methods must do) and to estimate the reward normalization (as REINFORCE algorithms must do) while at the same time is harmonizing the model with its test-time output ([81](#)).

## 2.5 The evolution of attention mechanisms

This section presents the evolution of the attention mechanisms used in Vision to Language tasks (e.g. image captioning). Recent neural REG approaches ([63](#), [67](#), [116](#)) have studied the effect of different visual features in the quality of the produced referring expressions. Specifically, such approaches extract the visual features from the last layers of a pretrained CNN and then those features are used as a conditioning element for the language model. The main advantage of this approach is that it extracts a compact representation of the whole context of image. However, the features extracted with this approach lack in granularity since all the salient objects or regions are compressed into a single vector. Furthermore, a global representation may result in sub-optimal solutions due to the fact that the probability of a word may be affected by noisy context or irrelevant parts of the image. The attention mechanism tries to overcome these problems by connecting the hidden states of the language model with a set of fine-grained visual features.

[Xu et al. \[109\]](#) introduced the first model that utilizes attention over the spatial



output grid of a convolutional layer. This mechanism enables the model to focus on different parts of the of the grid by selecting a subset of features relevant to each generated word. In more detail, the activation of the last convolution layer of the VGG network are extracted. Then the attention score is computed for each grid element. The attention score defines the relative importance of that part of the grid for the generation of the next word.

[Chen et al. \[9\]](#) proposes a model that considers both spatial and channel-wise attention to compute an attention score. They argued that the attention mechanism proposed by [Xu et al. \[109\]](#) loses spatial information gradually due to the fact that it calculates the attention for a subset of visual features and utilizes the activations only of the last layer. Therefore, they calculate the attention score from different channels and multiple layers of the CNN.

Another line of research investigates whether information about human gaze can be beneficial for language models. [Sugano and Bulling \[86\]](#) presents a new perspective on gaze-assisted image captioning by studying the interplay between human gaze and the attention mechanism. Specifically, their model integrates human gaze information input to the attention mechanism proposed by [Xu et al. \[109\]](#). The attention score for each image region is weighted based on whether they are fixated or not. Subsequent research replaces gaze information with saliency information predicted by human written descriptions ([78, 92](#)).

Human attention can be focused spontaneously by top-down signals that are guided by the objective of the task at hand (e.g., looking for something). The models described so far exploit attention mechanisms driven by task-specific context and thus can be seen as top down systems ([1](#)). In other words, a word is predicted while the language model attends a subset of a feature grid the geometry of whom is irrespective of the image content. However, little consideration is given to how the attended image regions are determined. To address this issue [Anderson et al. \[1\]](#) proposes the addition of a bottom-up path guided by an object detector responsible for proposing image regions. In more detail, the object detector detects objects in two stages. The first stage predicts object proposals by sliding over intermediate CNN features. The second stage employs

---

regions of interest pooling to extract a small feature map for each proposal. In order the object detector to predict a dense and rich set of detections, attribute classes alongside object classes are predicted. Although utilizing pooled vectors from image regions has been the de-facto attention approach in image captioning and has been followed by many models in the recent years [Huang et al. \[38\]](#) attention only fixes on a single visual region at each step. In order to enable the attention mechanism to attend multiple regions [Zha et al. \[122\]](#) introduces a sub-policy network that explicitly considers the previous visual attentions as context and decides whether the context is used for the current word/sentence generation given the current visual attention. Similarly [Pedersoli et al. \[76\]](#) proposes an attention mechanism that models the interplay between the RNN state, image regions, and word embeddings through pairwise interactions. The image-specific attention areas are proposed with the use of a spatial transformer. Specifically, they use a localization network to locally regress an affine transformation of each position of the feature map. Then, the feature vector for each region is bilinearly interpolated with respect to the anchor boxes.

To further improve the expressively of images region and their spatial relationships, another line of research constructs graphs over the detected objects in an image to better represent their spatial and semantic connections. The first to incorporate graphs is [Yao et al. \[112\]](#) that proposes to use a graph convolution network in order to integrate semantic and spatial object relationships into the image encoder. They build the semantic graphs by learning a semantic relation classifier that predicts either an action or the relationship between every two regions within the image. Furthermore, the spatial relationships of the objects are represented by the relative geometrical position of the bounding boxes of the object pairs. [Yang et al. \[111\]](#) proposes the Graph Auto-Encoder (SGAE) that exploits a graph based representation (i.e. scene graph) of both images and sentences. The scene graph is a directed graph that connects objects, their attributes, and their relations. [Yao et al. \[113\]](#) proposes a model that integrates the hierarchical structure of the image to the image encoder. The key idea of this method is to build a hierarchical tree, where the root is the whole image, the intermediate nodes represent image regions and the leafs represent detected objects in the regions. Then, a

Tree-LSTM (90) is employed to interpret the hierarchical structure and extract the final image encoding.

Although graph encoding enables the model to leverage spatial and semantic relationships between objects within an image, the manual construction of the graph structures limits the interactions between visual concepts. As a solution to this problem the self-attention is proposed by Vaswani et al. [94]. Within this framework each element of a set is connected with all the others and a refined representation of the set of elements can be computed through residual connections. The self-attention mechanism was originally proposed for machine translation by 94 and since has dominated many natural language processing fields. Yang, Zhang and Cai [110] is amongst the first attempts that utilize a self-attentive mechanism in order to encode semantic and spatial relationships between objects within the image. Li et al. [50] proposes a transformer based model that exploits semantic and visual information simultaneously through a semantic and visual encoder. Each encoder is build with self-attention and feed-forward layers. Then a gating mechanism regulates the fusion of the output of each encoder into the decoder. Herdade et al. [31] introduced a geometric self attention mechanism that explicitly accounts for the spatial relationships between detected objects within the image. Specifically, the attention weights are scaled through a geometric weight that is computed for every object pair. This idea is extended by Guo et al. [26] that proposes a self attention mechanism that dynamically computes the relative geometry relationships between objects in the image.

Huang et al. [39] proposes a new attention mechanism called “Attention on Attention”, where the attended regions are weighted by a context gate. Specifically, they first generate an information vector and an attention gate using the attention result and the current context, and the final attention score is computed by element-wise multiplication of those vectors.

## 2.6 Conclusions

This chapter presented an overview of related work to this thesis. Specifically, previous work on neural REG rely on incorporating contextual information by using visual features, appearance attributes (61), location features (116) and global image features as target object representation and complex architectural set-ups that utilize a comprehension module (61, 63, 115) to decrease the ambiguity of the produced referring expressions. However, they do not investigate the effect that the language model architecture, the decoding algorithm and the learning strategy has on the produced referring expressions. Furthermore, the aim of those systems is to reduce the ambiguity of the produced referring expressions, while other essential natural language attributes such as *diversity* and *naturalness* have received less attention. Therefore, this thesis explores how the choices of the language model architecture, the decoding algorithm and the learning strategy affect the produced referring expressions w.r.t diversity, ambiguity and naturalness.

Furthermore, in Section 2.4 it was discussed how RL provides a better solution than traditional methods. However, leveraging RL methods creates its own training challenges. First, in REG the action space is a high-dimensional discrete space that is massive comparing to the size of actions in a robotic or game-playing problems. Hence, sample efficiency and high variance are two of the main issues in applying RL in REG. Although REINFORCE-based models such as SCST (81) are preferred in most of the current approaches, sampling only one prediction is sample-inefficient. Therefore, this thesis will be concerned with the development of a REINFORCE-based model is sample-efficient.

## *Language Models*

---

Recent neural approaches in REG follow the encoder-decoder model architecture (8, 115, 116, 119). Originally the encoder-decoder architecture was proposed for machine translation (11), where the task is to translate a sentence written in a source language into a sentence to the target language. Within the machine translation framework, an encoder RNN encodes the source sentence into a fixed length vectorial representation which in turn is used to initialize the hidden state of a decoder RNN that is responsible to generate the translation. The first to adapt this elegant recipe into Vision to Language tasks were Vinyals et al. [102] that proposed to replace the encoder RNN with a deep convolutional neural network. They argued that a pretrained CNN in an image classification task is capable of producing rich image representations. Specifically, those representations are usually the output the output of last convolutional layers of a network. Many different CNNs has been used in the literature as image decoders, e.g. VGG (85), Resnet (28) or Inception (89). As a decoder, a plethora of early works used a vanilla RNN. While early approaches were effective, there is a number of limitations that stem from those approaches. First, older information tends to fade from the context as new information is integrated into the context. Second, the context is likely to be dominated by recent information so when an error is made it will result in a cascade of errors through the rest of the sequence. Third, the visual stimuli is static, i.e. does not change during the generation of a referring expression. The visual representation is only fed to the model at the start of the generation. However, as the generation progresses,

this paradigm may result in sub-optimal expressions due to noisy contexts or irrelevant regions while generating a specific word. Finally, such approaches describe the target object as a whole instead of focusing on fine-grained details of the target object that will reduce the ambiguity of the description.

Attention mechanisms can address the aforementioned limitations by dynamically focusing on different parts of the visual stimuli as referring expressions are generated. Such mechanisms are effective for a variety of sequential prediction tasks, such as storytelling (19, 36), summarization (25, 91), dialogue systems (55, 103), and image captioning (102, 108). In this thesis, a novel object attention model is proposed that allows the attention mechanism to be calculated at the level of the target object. The object attention mechanism aims at connecting the encoder and decoder, thus aligning better object visual features-to-word interactions.

Despite the success that RNN based attention models have achieved, there are two limitations that stem from such models. First, the attention mechanism only models the relationship between visual features and words, while neglecting the word-to-words interactions. Secondly, the LSTM based models are shallow, thus may fail to capture semantic attributes that are kindred to language. The transformer model was proposed to fill this gap, by simultaneously capturing the intra and inter modal interactions in a self-attention fashion through a deep stack of attention blocks. Transformers have revolutionized NLG fields such as machine translation, where the machine generated translations surpass the performance of those produced by human experts (94). Therefore, in order to further prove the benefits of attention in REG, this thesis investigates the effectiveness of the original architecture and proposes a novel architecture that follows a different layer configuration in order to provide the network with a global “context” signal by connecting each layer of the encoder with the respective layer of the decoder.

### **3.1 Contributions**

The contributions of this chapter are the following:

- A novel attention language model is proposed. The key novelty of this model is the incorporation of an object attention mechanism that assists the model in determining both the relationship between objects, but also determine fine appearance details of the target object. This is due to the fact that the proposed approach is able to consider all the information pertaining to an object simultaneously. It is shown that it leads to significant improvements over the non-attentive LSTM that has exclusively been used in recent works. This contribution is described by [Panagiaris, Hart and Gkatzia \[73\]](#).
- A novel transformer-based language model for REG that injects a form of context to the architecture by connecting each layer of the encoder to the respective decoder layer. The main benefit of the proposed approach is that it encodes the visual stimuli in a multi-level fashion. Consequently, both low and high level relationships are exploited. It is shown that it leads to significant improvements over the standard transformer architecture. This contribution is described by [73](#).

### 3.1.1 LSTM-based language model

This section presents a description of the standard LSTM-model that has been used in recent works in neural REG. Specifically, within this framework the representation of the target object is extracted with the use of a pre-trained CNN network and then this representation is embedded through a linear projection  $W_I$ . Each word  $x_t$  is represented as one-hot vector, mapped to the same space as the object representation through a linear embedding. The start of each sequence is denoted by a special **BOS** token, while the special stop token **EOS** denotes the end of the sequence. For the generation of the sequence of words, an LSTM model is used. The image features are only used as an input to  $t = 0$  in order to initialize the LSTM with visual features. Then, at each time step  $t$ , its output depends on the previously generated words and the hidden units, which encode the knowledge of the observed input up to this time step. More formally, the model is defined by the following update rules:

$$(3.1) \quad i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + b_i) \quad (\text{Input gate})$$

$$(3.2) \quad f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + b_f) \quad (\text{Forget gate})$$

$$(3.3) \quad o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + b_o) \quad (\text{output gate})$$

$$(3.4) \quad c_t = f_t \odot c_{t-1} + i_t \odot \sigma(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \quad (\text{memory cell})$$

$$(3.5) \quad m_t = o_t \odot \tanh(c_t) \quad (\text{hidden state})$$

$$(3.6) \quad p_{t+1} = \text{softmax}(m_t)$$

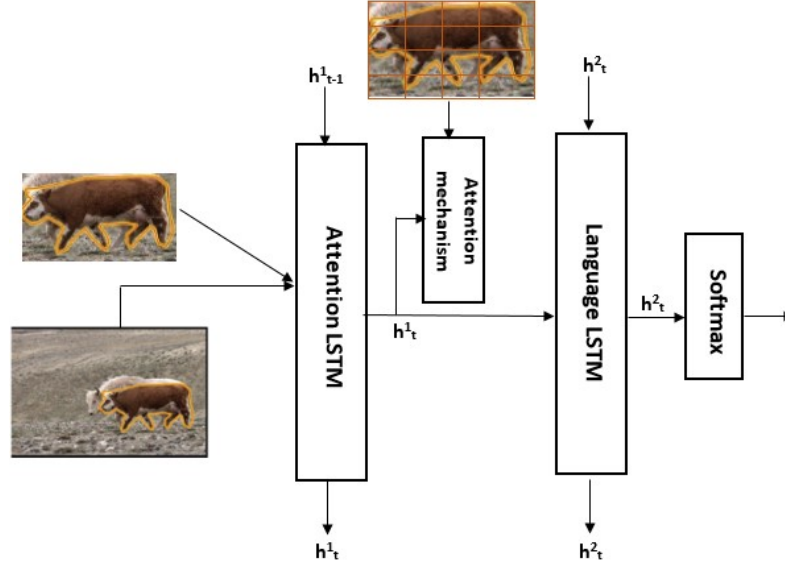
where  $\sigma$  is the sigmoid function and  $p_{t+1}$  is the probability distribution over all words. The  $W$ ,  $b$  matrices are learnable parameters and biases.

### 3.1.2 Object attention language model

Most neural REG approaches in literature are based on a translational approach, with a visual encoder and a linguistic decoder. A fundamental challenge in machine translation is that the generation of each word is not independent of the previous generated words. The generated words influence the meaning therefore the translation. This challenge is even more important when translating across modalities, i.e. from images to text, where the model has to decide how the target object should be uniquely described. A common solution for this challenge is the incorporation of attention mechanisms. For instance, recent models in image captioning try to solve where to look in the image during the encoding stage (1).

Attention models instead of utilizing a static visual representation, as the model described in Section 3.1.1, they dynamically re-weight the spatial visual features to “attent” on specific visual regions at each time step. The definition of spatial image features is generic. However, in this thesis the models are leveraging the spatial output grid of the last convolutional layer of a pre-trained CNN model as spatial visual features. Therefore, this thesis proposes a novel object attention mechanism for REG. Specifically, given a set spatial object features the proposed model uses an attention mechanism to weight each feature during the generation process. Conceptually, the model is





**Figure 3.1:** Overview of the proposed object attention language model.

composed of two LSTM layers, where the attention mechanism is implemented in the first layer and the second layer plays the role of the language model and follows the update rules described in Section 3.1.1. The overall architecture is depicted in Figure 3.1.

The input vector to the LSTM attention layer at each time step is comprised of the output of the language LSTM concatenated with the mean-pooled object features. Formally, the input to the LSTM attention layer is the following:

$$(3.7) \quad v_i = [r, \bar{o}, \bar{I}, h_{t-1}^L]$$

where  $\bar{o}$  is the concatenation of the mean-pooled object region features (i.e.  $\bar{o} = \frac{1}{k} \sum_i o_i$ );  $r, I$  are the CNN extracted features for the target object and image respectively and  $h_{t-1}^L$  is the previous hidden state of the language LSTM. It is assumed that this input representation is expressive enough for the context of the image and the state of language model in order to steer the model to information that is important for the target object.

The attention weighted annotation vector  $a_{i,t}$  for the uniform grid of the object region is computed as follows:

$$(3.8) \quad a_{i,t} = \mathbf{w}_a^T \tanh(W_{oa}v_i + W_{ha}h_t^1)$$

$$(3.9) \quad \boldsymbol{\alpha}_t = \text{softmax}(\mathbf{a}_t)$$

where  $W_{oa}v_i \in \mathbb{R}^{A \times D}$ ,  $\mathbf{w}_a \in \mathbb{R}^A$  and  $W_{ha} \in \mathbb{R}^{A \times d}$  are learnable parameters and  $A$  indicates the dimensions of the attention layer. Finally, the attention derived *object* visual features that will be used as input to the language LSTM are given by:

$$(3.10) \quad \hat{o}_t = \sum_{i=1}^K \alpha_{i,t} o_i$$

Specifically, the input to the language LSTM is the combination of the attended object features and the hidden state of the attention LSTM  $h_t^a$ . Formally the input  $i_t^l$  is the following:

$$(3.11) \quad i_t^l = [\hat{o}_t, h_t^a]$$

Using the notation  $y_{1:T}$  to refer to a sequence of words  $(y_1, \dots, y_T)$ , at each time step  $t$  the conditional distribution over possible output words is given by:

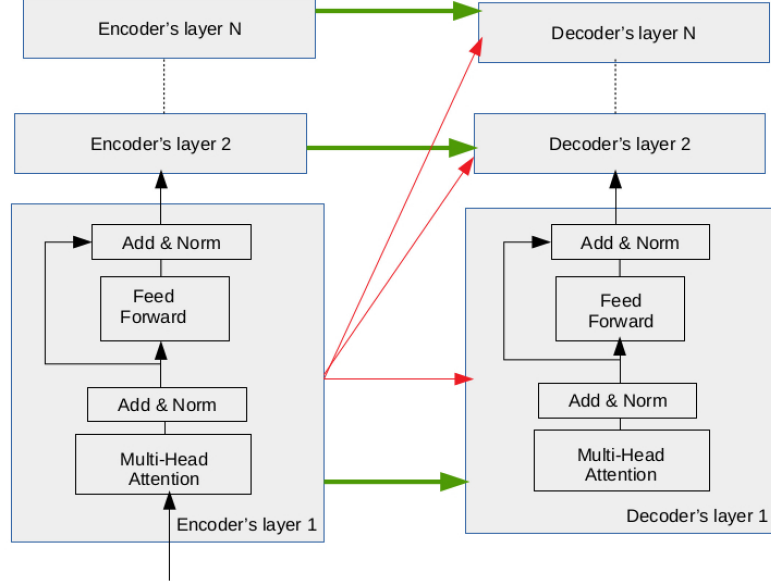
$$(3.12) \quad p(x_t | x_{1:t-1}) = \text{softmax}(W_p x_{1:t-1}^2 + p)$$

where  $W_p \in \mathbb{R}^{|\Sigma| \times M}$  and  $p \in \mathbb{R}^{|\Sigma|}$  are learned weights and biases. The distribution of a complete output sequences is the product of the conditional distributions:

$$(3.13) \quad p(x_{1:T}) = \prod_{t=1}^T p(x_t | x_{1:t-1})$$

### 3.1.3 Transformer

The transformer model can be conceptually divided into an image encoder and a decoder module. The encoder learns in a self-attention fashion visual representations, while the decoder makes use of the attention-derived visual representations to generate the output. In order to handle variable-length inputs, such as image regions and text



**Figure 3.2:** Overview of the transformer architecture (94). The red arrows illustrate the original connectivity between the encoder and decoder, while the green arrows illustrate the proposed connectivity.

sequences, the transformer employs two attention mechanisms: (1) the scaled dot-product attention; and (2) the multi-head attention. The former type of attention is first introduced since it is the most important function of the transformer model.

The scale-dot product function receives as an input: a query  $q \in \mathbb{R}^d$ , a set of keys  $k_t \in \mathbb{R}^d$  and values  $v_t \in \mathbb{R}^d$ , where  $t \in \{1, 2, \dots, n\}$ . It outputs the weighted sum of value vectors  $v_t$ . For practical reasons, all the keys and values are packed into matrices  $K = [k_1, \dots, k_n] \in \mathbb{R}^{n \times d}$  and  $V = [v_1, \dots, v_n] \in \mathbb{R}^{n \times d}$  respectively. More formally, given a set of queries  $Q = [q_1, \dots, q_m] \in \mathbb{R}^{m \times d}$  the scaled dot-product attention operator is defined by:

$$(3.14) \quad \text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where  $d$  is a scaling factor. In this thesis the implementation of Vaswani et al. [94] was followed and a scaling factor of  $d = 64$  is used. The scaling factor indicates the cardinality of the value, key, and queries vectors. In order to attend different representation sub-spaces, the multi-head attention is introduced. It consists of  $h$  independent scaled dot-product operators named as “heads”. Each attention head first calculates the queries, keys, and values that are projected into  $h$  sub-spaces as follows:

$$(3.15) \quad \text{MultiHead}(Q, K, V) = \text{Concat}(h_1, \dots, h_h)W^o$$

$$(3.16) \quad H_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

where  $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d \times d_h}$  are the projection matrices for the  $h$  independent heads, while  $W^O \in \mathbb{R}^{h \times d_h \times d}$  is the output projection matrix that aggregates the information from  $h$  heads. In this thesis the optimal number of heads for a REG model was explored. In particular a number of models were trained ranging the number of heads from one to twenty. Hence, empirically was found that the optimal number of heads is eight. Therefore, all of the transformer-based architectures presented in this work employ eight heads.

The transformer leverages stacks of identical layers to mimic the encoder-decoder architecture. The overall architecture of a transformer-based model, is illustrated in Figure 3.2. Specifically, the encoder is a stack of  $N$  identical layers. Each layer is comprised of a multi-head attention mechanism given by the Equation 3.15. The second component is a position-wise feed-forward network that is applied to the output of the multi-head attention layer as follows:

$$(3.17) \quad \text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

where  $W_1, b_1, W_2, b_2$  are the weights and biases of the two fully connected layers. Finally, residual-connections (30) that are followed by layer-normalization (3) are applied to the outputs of the self-attention and the feed-forward layer. The decoder's first layer receives as input the output of encoder's last layer. Similarly to the encoder, the decoder is a stack of  $N$  identical layers. However, in addition to the two sub-layers in each encoder layers, a third module is added to the decoder layers to perform multi-head attention over the encoder's output.

The transformer architecture was originally designed and proposed for automatic text translation. In automatic text translation, a word is either to the left or right of

another word with different distances. However, images are two-dimensional, and thus the relative relationship between images regions has a larger degree of freedom than the semantic units of a sentence. In other words, an image region can contain or be contained in another region. Secondly, in machine translation a word is mapped to a word of another language (one to one decoding), whereas for an image or an image region its content can be described with different ways given that its content, attributes and relationships are depended with other regions of the image (one to many decoding).

To overcome the aforementioned shortcomings, this thesis proposes a modification of the transformer’s internal architecture. Specifically, a different connectivity pattern between the encoder and the decoder is proposed. Each layer of the encoder is connected with the respective decoding layer. The proposed connectivity is illustrated in Figure 3.2 with the green arrows, while the original connectivity is shown with red arrows. Specifically, in order for a word to be predicted there should be a form of visual information that influences the likelihood. The original configuration utilizes a fixed representation throughout the network. However, fixed visual representations might be unable to capture the transitioning dynamics between the visual focus and words. Therefore, visual features with different degrees of modification are incorporated at each layer, to better model the interdependencies of different visual elements and words. Finally, a number of different connectivity patterns were explored. Specifically, connecting every layer with every layer of the decoder resulted in a reduction of 0.17 in CIDEr scores compared to the original configuration. However, combining the output of each layer and then feed it in each layer of the decoder resulted in an increase of 0.009 CIDEr score compared to the original architecture.

### 3.1.4 Training objective

Let  $\theta$  denote the parameters of the language models described in Sections 3.1.1, 3.1.2 and 3.1.3. Let  $\{x_1^*, x_2^*, \dots, x_T^*\}$  be a ground-truth referring expression, the model parameters  $\theta$  are trained to minimize the cross entropy loss as follows:

$$(3.18) \quad L(\theta) = - \sum_{t=1}^T \log(\pi_{\theta}(x_t^* | x_{1:t-1}^*, I, r))$$

where  $\pi_{\theta}(x_t|x_{1:t-1}, I, r)$  is the probability distribution of the token  $x_t$  given all the previous generated tokens  $\{x_1, x_2, \dots, x_{t-1}\}$  and the visual features  $I, r$ .  $T$  denotes the length of the sequence.

## 3.2 Experimental design and results

### 3.2.1 Implementation Details

**Visual Features** The visual representation that was used is a 4096-dimensional vector that is a concatenation of: (1) a 2048-dimensional vector of the target object region; (2) a 2048-dimensional vector representation of the whole image that serves as context features. As main feature extractor we used ResNet-152 (29). In more detail, for the object region features, the aspect ratio of the region was kept constant and was scaled to  $224 \times 224$  resolution. The margins were padded with the mean pixel value, following (66). The attention features were extracted as follows. First, each target region was encoded with the final convolutional layer of ResNet-152. Then, bilinear interpolation was applied to resize the output to a fixed size representation of  $7 \times 7$ ,  $10 \times 10$  and  $14 \times 14$ . However, empirically was found that the  $14 \times 14$  performs best. Both the object region and image features are pre-extracted and no fine-tuning was performed. The input visual representation was kept fixed across all the experiments.

**Training** For the best performing LSTM and LSTM+ATT, the dimensions of the LSTM’s hidden state, image feature embeddings, and word embeddings was set to 512. The batch size is set to 128 objects. The learning rate is initialized to be  $5 \times 10^{-4}$ , and decays by a factor of 0.8 every three epochs.

The best transformer model consists of 3 fully connected encoding and decoding layers. The dimensionality of each layer was set to 512 and 8 attention-heads were used. Every feed-forward layer is followed by a dropout with a rate of 0.1. The learning rate is initialized to be  $5 \times 10^{-4}$  and decays by a factor of 0.8 every three epochs, with 20000 warmup steps. The batch size was set to 10 objects.

### 3.2.2 Evaluation

**Evaluation of one-shot referring expressions:** For the evaluation of the models presented in this thesis both intrinsic and extrinsic evaluation is performed. For the former, standard automatic metrics that have been used in REG (67, 116, 119) are used. Automatic metrics compare the generated referring expression with the human ones. The models are evaluated with the use of the following metrics:

- **BLEU:** It is a precision-based metric that computes the n-gram overlap between the machine generated text and the human written text. BLEU is the ratio of the number of overlapping n-grams to the total number of n-grams in the human written text. In other words, it measures the similarity of a machine generated text with the human generated text. Given the short length of referring expressions the models are evaluated on  $BLEU_1$  for uni-grams.
- **CIDEr:** It was first proposed in the context of image captioning where each image is accompanied by multiple human written descriptions. The underlying idea of CIDEr is that n-grams that are relevant to an image would occur frequently in its set of human written descriptions captions. Therefore, it weighs every n-gram in a sentence based on its frequency in the corpus and in the set of human written descriptions for a particular image, using TF-IDF (term-frequency and inverse-document-frequency). Furthermore, for n-grams that appear frequently in the entire corpus of human written descriptions a lower weight is assigned.

Previous work has shown that automatic evaluation metrics do not correlate well with human judgments (44, 115, 116, 120). Unlike other generation tasks such as image captioning, here a referring expression is successful if it describes the target object unambiguously. Thus, human evaluation was conducted in order to measure the task success of the systems. Specifically, all human evaluation experiments were conducted at Amazon Mechanical Turk. An Amazon Mechanical Turk worker was able to participate in this human evaluation experiment only if the worker was a native speaker of the English language, was located in English-speaking countries, had an

approval rate of 99% and had successfully completed 1000 tasks. Each worker was only allowed to participate in a task only once. In each task the workers were presented with an image and an expression and were asked to draw a box around the object that they believe is best described by that expression. If two workers chose the correct object, then the expression was considered successful. For each test set 60 objects were randomly selected. For each test set 180 unique ratings were collected, at a cost of 1 USD per rating. In addition to the evaluation of the success of referring expressions, annotators were asked to rate (in a typical Likert scale from 1 to 5) the statements below following [Mitchell et al. \[69\]](#):

- **Q1-Grammaticality:** The description is grammatically correct.
- **Q2-Main aspects:** The description does not describe the main attributes correctly.
- **Q3-Correctness:** This description does not include extraneous or incorrect information.
- **Q4-Naturalness:** It sounds like a person wrote that description (Yes/No)

### 3.2.3 Datasets

The models presented in this thesis are trained on RefCOCO and RefCOCO+ (116) which are built on MSCOCO dataset (58). Specifically, the collection of the expressions of RefCOCO(+) was based on the ReferIt Game (43), an interactive setting where two players alternate between two roles: (1) speaker: generating referring expressions; (2) listener: identifying the described object within an image. The RefCOCO(+) images contain on average 3.9 objects of the same category within an image and they contain approximately 150k referring expressions for 50k objects. Although both datasets contain similar images, the referring expressions for each dataset are quite different due to different data collection instructions. In particular, for RefCOCO+, the use of absolute location words (e.g. top right, bottom left, etc.) was not allowed and thus the RE are *appearance* focused, while for the RefCOCO the use of *location* is essential





**Figure 3.3:** Human written referring expression for one target object (green box) for each of the test sets of RefCOCO and RefCOCO+.

in order for the target object to be successfully individualized. Furthermore, for each dataset different test splits are provided. The predefined test splits for both datasets are divided between person vs object splits. In particular, images containing people are in “TestA” and images that contain all other object categories are in “TestB”.

### 3.2.4 Attention-based REG results

In order to demonstrate the advantages of the proposed object attention, a detailed comparison between the attention model and the standard LSTM was performed. Furthermore, in order to determine whether the difference caused by incorporating the object attention was statistically significant a two-tailed t-test with paired samples was performed. The results for the two considered datasets are shown in Table 3.1. Specifically, the proposed attention model results in higher scores than the standard LSTM. The difference in scores was found statistically significant (using a significance level  $\alpha = 0.05$ ). The significant improvements in CIDEr and  $BLEU_1$  are in line with the assumption made in this thesis that adding the object attention mechanism would assist

		RefCOCO				RefCOCO+			
		testA		testB		testA+		testB+	
Model Type	Decoding Method	$BLEU_1$	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$	CIDEr
LSTM	Greedy	0.490	0.762	0.523	1.332	0.444	0.633	0.373	0.710
	Beam	0.477	0.758	0.510	1.340	0.429	0.656	0.384	0.837
LSTM +ATT	Greedy	<b>0.594</b>	<b>1.033</b>	<b>0.609</b>	1.552	0.512	0.884	0.424	0.858
	Beam	0.577	1.013	0.599	<b>1.573</b>	<b>0.491</b>	<b>0.881</b>	<b>0.424</b>	<b>0.857</b>
p-value		<b>&lt;0.01</b>	<b>&lt;0.01</b>	<b>&lt;0.01</b>	<b>&lt;0.01</b>	<b>&lt;0.01</b>	<b>&lt;0.01</b>	<b>&lt;0.01</b>	<b>&lt;0.01</b>

**Table 3.1:** Comparison of different automatic metrics for the attention model (denoted as “LSTM-ATT”) and the standard LSTM model. The proposed attention model results in significantly higher CIDEr and  $BLEU_1$  scores in both datasets. The p-values are the result of two-tailed t-tests using paired samples.

the model in determining both the relationship between objects, but also determine fine appearance details of the target object. This is due to the fact that the proposed approach is able to consider all the information pertaining to an object simultaneously.

To illustrate the advantages of the proposed approach, examples of objects with the corresponding referring expressions generated by each model are presented (see Figure 3.4 and Figure 3.5). The referring expressions presented here were generated using the following steps: both models were trained with MLE and were greedily decoded. For the chosen examples there was a significant improvement between the CIDEr scores of the expressions generated by the attention model and those generated by the standard LSTM. The collection of objects and expressions for RefCOCO and RefCOCO+ is shown in Figure 3.4 and Figure 3.5 respectively. It should be noted that, during the collection of RefCOCO dataset, no restrictions were placed on the type of language that can be used in the referring expressions, while in RefCOCO+ dataset location words were not allowed. Thus, this dataset contains referring expressions that are based on appearance attributes. Specifically, the images in testA that are presented in Figure 3.4, illustrate an improvement in determining when a relationship between objects should be expressed, as well as in determining what that relationship should be. In addition, the images in testB presented in Figure 3.4, illustrate an improvement in including appearance and location attributes. The improvement in including appearance attributes can be further noticed in the referring expressions of RefCOCO+ dataset presented in Figure 3.5.



**Figure 3.4:** Examples of objects and expressions drawn from RefCOCO dataset, for which the CIDEr scores of the attention model show an improvement over the standard LSTM. The target object is highlighted with a red box.



**Figure 3.5:** Examples of objects and expressions drawn from RefCOCO+ dataset, for which the CIDEr scores of the attention model show an improvement over the standard LSTM. The target object is highlighted with a red box.

### 3.2.5 Transformer-based REG results

Table 3.2 shows the results for the ablation study regarding the transformer model discussed in Section 3.1.4. The original configuration (denoted as Transformer 6 in Table 3.2) of the transformer (94) serves as the baseline. To determine whether the

Model	RefCOCO				RefCOCO+			
	testA		testB		testA		testB	
	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$
Transformer 6	0.837	0.506	1.340	0.546	0.772	0.460	0.763	0.387
Transformer 6 (OURS)	0.852	0.513	1.355	0.552	0.798	0.467	0.791	0.395
Transformer 3	0.922	0.524	1.442	0.581	0.911	0.515	0.894	0.412
Transformer 3 (OURS)	<b>0.938</b>	<b>0.586</b>	<b>1.464</b>	<b>0.586</b>	<b>0.938</b>	<b>0.529</b>	<b>0.913</b>	<b>0.424</b>

**Table 3.2:** The impact of depth in the performance of the transformer model. Transformer 6 and 3 indicate that the decoder and the decoder consist of 6 and 3 layers respectively. The layer configuration follows the one proposed by 94. “Ours” indicates that each layer of the encoder is connected with the respective layer of the decoder.

changes in the configuration of the model result in statistically significant differences for each of the considered metrics, we performed a two-tailed t-test with paired samples as described in Section 3.2.4.

First the effect of the number of layers is investigated. It is hypothesised that, given the model was initially proposed for machine translation, a task with considerable longer sentences than REG and larger training sets, a shallower architecture might result in better performance. Table 3.2 shows that reducing the depth (Transformer 3 in Table 3.2) of the network leads to considerable improvements in both  $BLEU_1$  and CIDEr scores. For instance, in RefCOCO+ testA, decreasing the number of layers leads to an improvement from 0.772 to 0.911 in CIDEr values. The score difference was statistically significant (using a significance level  $\alpha = 0.05$ ). However, decreasing the number of layers to less than three results in significant reduction in CIDEr scores. Specifically, a two layer network achieves CIDEr scores of 43.4 and 0.37 for RefCOCO testA and testB respectively. The same trend applies when the number of layers increases. Specifically, a eight layers network achieves scores 23.26 and 25.05 for RefCOCO testA and testB respectively.

The effect of connecting each layer of the encoder to the respective layer of decoder is investigated. The results are shown in Table 3.2, where “Transformer (OURS)” stands for the proposed model. Specifically, for all of the considered metrics, the proposed transformer produces higher scores than the standard transformer.

Examples of generated REs are illustrated in Figure 3.6. The referring expressions



**Figure 3.6:** Examples of objects and expressions drawn from both RefCOCO and RefCOCO+ datasets, for which the CIDEr score for the proposed transformer model show an improvement over the standard transformer. The target object is highlighted with a red box.

presented here were generated using the following steps: both models were trained with MLE and were decoded using greedy decoding. In all images presented in Figure 3.6, we observe that the proposed model improves over the standard transformer in inferring fine appearance (e.g. “number 29” top left image in Figure 3.6) and location attributes of the target object. This is in line with the expectation that utilizing features with different degrees of modification at each layer, will better model the interdependencies of different visual elements and words.

### 3.2.6 Comparison between the proposed language models

This section compares the two proposed models. Specifically, the results presented in Table 3.3 demonstrate that the proposed transformer model is more effective in task success compared to the attention model. Additionally, it is observed that it scores higher in naturalness of the produced referring expressions across datasets. Furthermore, the referring expressions generated by the proposed model were describing the main aspects of the target correctly as was judged by the human annotators. Finally, the proposed transformer model achieves state-of-the-art results in both datasets.

RefCOCO testA					
	Task success	Naturalness	Grammaticality	Main Aspects	Correctness
LSTM+ATT	71.66%	92.85%	4 (3.41, 0.68)	2 (2.23, 0.92)	3 (3.30, 0.77)
Trasnformer	<b>78.33%</b>	<b>96.42%</b>	4 (3.69, 0.62)	3 (2.87, 0.88)	4 (3.64, 0.88)
Best by Yu et al. [115]	76.95%	-	-	-	-

RefCOCO testB					
	Task success	Naturalness	Grammaticality	Main Aspects	Correctness
LSTM+ATT	66.66%	<b>98.92%</b>	4 (3.71, 0.83)	2 (2.23, 0.92)	3 (3.30,0.77)
Transformer	73.33%	98.21%	3 (2.78, 0.61)	3 (2.28, 0.79)	3 (2.85, 0.71)
Best by 115	<b>78.10%</b>	-	-	-	-

RefCOCO+ testA					
	Task success	Naturalness	Grammaticality	Main Aspects	Correctness
LSTM+ATT	76.66%	93.64%	4 (4.12, 0.92)	2 (1.91, 0.93)	3 (3.25, 0.96)
Transformer	<b>80.00%</b>	<b>95.44%</b>	4 (3.82, 0.38)	3 (2.11, 0.68 )	4 (3.78, 0.55)
Best by 115	58.85%	-	-	-	-

RefCOCO+ testB					
	Task success	Naturalness	Grammaticality	Main Aspects	Correctness
LSTM+ATT	55.00%	71.42%	4 (4.17, 0.92)	2 (1.91, 0.93)	3 (3.25, 0.96)
Transformer	<b>58.33%</b>	<b>92.85%</b>	4 (3.62, 0.51)	3 (2.89, 0.64 )	3 (3.07, 0.59)
Best by 115	58.20%	-	-	-	-

**Table 3.3:** Human Evaluation results. Median scores for systems, mean and standard deviation in parentheses.

### 3.3 Conclusions

In this chapter, the two proposed language models were presented. First, the benefits of incorporating an object attention mechanism in the language model were presented. Specifically, the proposed language models allow the attention mechanism to be calculated at the level of the referent object. It was demonstrated that applying this approach to REG, results in significant benefits compared to the standard LSTM model. The results on RefCOCO and RefCOCO+ show an increase, on average, of 0.26 and 0.12 in CIDEr scores respectively. The qualitative analysis showed that the attention mechanism results in an improvement in determining fine appearance attributes of the target object as well as an improvement in expressing the absolute and relative location of the target object. Unlike the standard LSTM, the proposed attention mechanism allows the language model to consider all the information pertaining the referent object at once. In other words, instead of letting the language model to hallucinate over the

attributes of the target object, the attention mechanism enables the language model to take multiple glimpses of the salient parts of the object’s region during generation. Furthermore, the human evaluation study showed that the proposed model performs comparable with the state-of-the-art (115). In particular, in RefCOCO+ testA it achieves an increase from 58.85% to 76.66% in task success.

Furthermore, to demonstrate the benefits of attention in REG, a transformer architecture was proposed that is noticeably effective in REG. It was shown that reducing the depth of the network from 6 layers to 3 results in an improvement in automatic metrics. The key novelty of this model is different connectivity pattern between the encoder and the decoder, by connecting each layer of the encoder with the respective decoder layer. The results on RefCOCO and RefCOCO+ datasets demonstrate significant improvements over the standard architecture. Moreover, the qualitative analysis showed how the proposed connectivity improves the spatial awareness and the inference of fine appearance attributes. Lastly, the human evaluation study (that follows the protocol described in Section 3.2.2 ) showed that the proposed transformer produces expressions that are more human-like, accurate and describing the main aspects of the target object better than the proposed attention LSTM. In addition, our results in task success improves over the state-of-the-art results in RefCOCO testA from 76.95% to 78.33% and in RefCOCO+ testA from 58.85% to 80.00%.

## *Sequence Level training objectives for REG*

---

The encoder-decoder model is typically trained to maximize the likelihood of a word given the history of generated words so far (see Section 3.1.4). This training approach is referred to as “Teacher-Forcing” (4). Although intuitive to train a model on token-level, during generation a model is evaluated based on its ability to optimize towards sequence level metrics resulting in a discrepancy between training and testing objectives. A second problem that stems from “Teacher-Forcing” is that during training, the model uses the ground-truth words to predict the next one, while during testing it uses its own predictions. This mismatch, coined as exposure bias (79), results in error accumulation during generation.

There is a large body of work that proposes solutions to the aforementioned exposure bias. Those approaches utilize reinforcement learning techniques (88). For example, [Ranzato et al. \[79\]](#) propose the use of the REINFORCE algorithm to directly optimize the non-differential evaluation metrics. A major limitation that stems from this method is that, the expected gradient exhibits high variance and without careful normalization is often unstable (81). An extension to the REINFORCE algorithm includes the bias correction with learned “baselines” (83, 118). [Rennie et al. \[81\]](#) propose an alternative way to normalize the reward. Specifically, they propose the self-critical sequence training (SCST), where instead of approximating the reward signal with learned “baselines”, it uses the output of the current model at test-time to calibrate the observed reward. A limitation of this approach is that it utilizes only one sample per data point that might be



insufficiently expressive for an observation. As a result, samples that poorly describe the observation will be heavily penalized, pushing the model to cover only high-probability zones. To minimize this effect, in this chapter is proposed a novel way to calculate the baseline of the REINFORCE algorithm. The proposed approach normalizes the reward by averaging over multiple-samples per observation.

Beside the high variance of the gradient there is a number of limitations that stems from RL training: (1) lack of per-token advantage, i.e. the REINFORCE algorithm makes the assumption that every token contributes equally to the whole sequence (107); and (2) reward configuration (79). Furthermore, effectively applying RL to REG has not been explored, with the exception of Yu et al. [115] who incorporate an additional module to reward discriminative REs by updating the speaker with a policy gradient algorithm. However, little is reported of how the RL was configured. Therefore, this chapter thoroughly investigates how to effectively train REG models with RL.

However, optimizing a model with RL leads to repetition of the produced output. To address the lack of diversity first this chapter proposes to combine RL objective with MLE. Although, this combined objective improves the output there is a considerable gap between MLE and RL methods w.r.t. to diversity. Therefore, this chapter proposes the use of minimum risk training (MRT) (84) as an alternative way of optimizing REG systems on sequence level. It further proposes a novel objective that combines both the MRT and MLE objective.

## 4.1 Contributions

The contributions of this chapter are the following:

- A novel optimization approach to REG based on the REINFORCE algorithm, that utilizes multiple samples per input to construct the baseline, rather than estimating the reward based on one sample. It was found that the proposed RL objective reduces the variance of the gradient compared to SCST training. This contribution is described by Panagiaris, Hart and Gkatzia [73].

- A novel strategy for training REG models using minimum risk training. It was shown that the proposed approach outperforms RL approaches w.r.t naturalness and diversity of the output. This contribution is described by 74
- As a complementary contribution, extensive analysis and benchmarking of RL training strategies for REG was conducted by exploring how different aspects such as the reward and the baseline reward configuration affect REG models. This contribution is described by 74.

## 4.2 Training REG with Reinforcement Learning

The main criticism of the MLE objective is that it does not take into consideration a task specific reward such as CIDEr. Secondly, it does not introduce a ranking amongst incorrect output referring expressions, since incorrect referring expressions are never considered during training. Therefore, the model will not learn to be robust to error accumulation during test time, since during training it never observed its own prediction.

As mentioned previously, reinforcement learning is used to bridge the gap between training and generation, by directly optimizing evaluation metrics (e.g., CIDEr) during training. The generation process can be cast into a reinforcement learning process as first described by Ranzato et al. [79]. In the classical reinforcement learning paradigm, the goal of an agent is to maximize the expectation of the reward  $r_t$  it receives for each action  $\hat{y}_t$  when interacting with its environment. More formally, an agent aims to maximize the following objective:

$$(4.1) \quad \mathbb{E}_{\hat{y}_1, \dots, \hat{y}_T \sim \pi_\theta(\hat{y}_1, \dots, \hat{y}_T)} [r(\hat{y}_1, \dots, \hat{y}_T)]$$

where  $\hat{y}_t$  is the word (i.e. action) sampled by the model at time  $t$  and  $r(\hat{y}_1, \dots, \hat{y}_T)$  is the observed reward for the actions  $\hat{y}_1, \dots, \hat{y}_T$ . Each agent performs an action under a specific policy  $\pi_\theta$ . The nature of the policy is application dependent. In the context of REG, the parameters of the agent (i.e. language model) define a policy. The agent selects an action, which is a candidate token from the vocabulary under the policy, until it generates the special token that denotes the end of the sequence. Once the agent

reaches the end of the sequence, it compares the sequence of actions under the current policy  $\hat{y}$  against the ground-truth sequence  $y$  and calculates a reward based on any task specific metric (e.g. CIDEr). The goal of the training is to parameterize the agent in order to maximize the reward. Formally:

$$(4.2) \quad \mathcal{L}_\theta = -\mathbb{E}_{\hat{y}_1, \dots, \hat{y}_T \sim \pi_\theta(\hat{y}_1, \dots, \hat{y}_T)} [r(\hat{y}_1, \dots, \hat{y}_T)]$$

In practice, however, the expected gradient is computed with only one sample acquired from the policy  $\pi_\theta$  as follows:

$$(4.3) \quad \nabla_\theta \mathcal{L}_\theta = -\mathbb{E}_{\hat{y}_1, \dots, \hat{y}_T \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(\hat{y}_1, \dots, \hat{y}_T) r(\hat{y}_1, \dots, \hat{y}_T)]$$

Applying the chain rule we have:

$$(4.4) \quad \nabla_\theta \mathcal{L}_\theta = \frac{\partial \mathcal{L}_\theta}{\partial \theta} = \sum_t \frac{\partial \mathcal{L}_\theta}{\partial o_t} \frac{\partial o_t}{\partial \theta}$$

where  $o_t$  indicates the input to the softmax function. Thus, the estimate of the gradient  $\mathcal{L}_\theta$  with respect to  $o_t$  is given by (118):

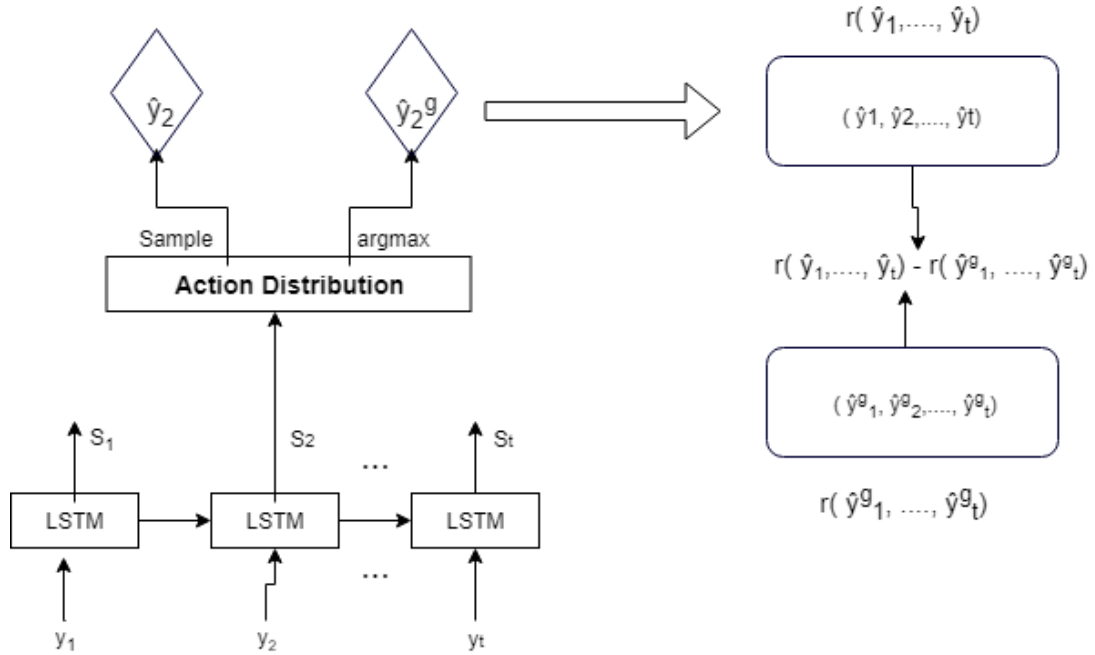
$$(4.5) \quad \frac{\partial \mathcal{L}_\theta}{\partial o_t} = \left( \pi_\theta(y_t | \hat{y}_{t-1}, h_t) - \mathbf{1}(\hat{y}_t) \right) (r(\hat{y}_1, \dots, \hat{y}_T) - r_b)$$

where  $r_b$  is a baseline reward. The role of the baseline is to guide the model towards actions with a reward  $r > r_b$  and penalize those that have a reward  $r < r_b$ . Furthermore, subtracting a quantity from the learning signal leads to lower variance, since it reduces its magnitude. This transformation leaves the gradient estimator unbiased because the baseline is a quantity that has zero expectation under the policy, since in this case:

$$(4.6) \quad \mathbb{E}_{\hat{y}_1, \dots, \hat{y}_T \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(\hat{y}_1, \dots, \hat{y}_T) r_b] = r_b \sum_{\hat{y}_1, \dots, \hat{y}_T} \nabla_\theta \pi_\theta(\hat{y}_1, \dots, \hat{y}_T)$$

$$= r_b \nabla_\theta \sum_{\hat{y}_1, \dots, \hat{y}_T} \pi_\theta(\hat{y}_1, \dots, \hat{y}_T) = r_b \nabla_\theta 1 = 0$$

This shows that the subtraction of the baseline leaves the gradient estimator unbiased. This algorithm has been coined in literature as the REINFORCE algorithm with



**Figure 4.1:** An illustration of the self-critical sequence training approach. In particular, a specific action  $\hat{y}_2$  is sampled and the greedy action  $\hat{y}_2^g$  is extracted. The difference of the rewards from sampling and greedy sequence is used to update the loss function.

a baseline (106). The reward, for example, could be calculated as the mean of the  $N$  rewards that are observed.

### 4.2.1 Self-critical sequence training (SCST)

An alternative way of reducing the variance was proposed by Rennie et al. [81]. In SCST, the reward is obtained by applying greedy search, the inference algorithm that is used at test-time. Thus, the following REINFORCE estimator is obtained:

$$(4.7) \quad \mathcal{L}_\theta = \frac{1}{N} \sum_{i=1}^N \log \pi_\theta(\hat{y}_i) \left( r(\hat{y}_{i,1}, \dots, \hat{y}_{i,T}) - r(\hat{y}_{i,1}^g, \dots, \hat{y}_{i,T}^g) \right)$$

where  $\hat{y}_{i,t}^g$  is an action sampled with greedy decoding. The central idea of the self-critical sequence training approach (see Figure 4.1) is to utilise the current model under the inference algorithm used at the test is used as baseline. Hence, samples from the model that return higher reward than  $\hat{y}$  will be positively rewarded, or increased in probability, while samples which result in lower reward will be suppressed. The difference of the rewards from sampling and greedy sequence is used to update the loss function. In practice, however, only a single sample is used to compute the expectation. From a classic RL point of view, using a single sample is a reasonable strategy, since it

might not be feasible to score multiple sampled actions for a state. However, from a data point of view, this is inefficient. Specifically, multiple samples can be evaluated without additional computational load. Secondly, optimizing a powerful model using one sample might have detrimental effects on its capacity. The assumption that one sample is sufficiently expressive does not always hold. As mentioned before, samples with higher rewards will be favored, while heavily penalizing samples that explain the observation poorly, leading to a lower bound of the likelihood. Therefore, the model will cover only the high-probability zones. An intuitive way to limit this crippling effect *is to average over multiple samples per data point*. The use of multiple samples per data point, provides sophisticated information leading to the construction of a more robust local baseline. Thus, this chapter proposes to use the REINFORCE with multiple-samples per input. A similar strategy, has been used on the travelling salesman problem presented by [Kool, Hoof and Welling \[47\]](#) where they use REINFORCE without replacement and in variational inference presented by [Mnih and Rezende \[70\]](#).

#### 4.2.2 REINFORCE with multiple-samples per data point:

Granted that the samples within the set are independent, a baseline  $b$  for the  $i$ -th item of a set can be constructed by averaging over the rest samples. In this thesis, both the arithmetic mean and the geometric mean  $b_i = \frac{1}{K-1} \sum_{i \neq j} r(\hat{y}^j)$  was used and a slight superiority of the latter was found. Therefore, the estimator in the Equation 4.8 becomes:

$$(4.8) \quad \mathcal{L}_\theta = \frac{1}{N} \sum_{i=1}^N \log \pi_\theta(\hat{y}_i) \left( r(\hat{y}_i) - \frac{1}{K-1} \sum_{i \neq j} r(\hat{y}_j) \right)$$

One of the advantages of the self-critical sequence training is that the baseline is based on the inference algorithm that is used at test-time without having to train an additional “critic” network. In particular, the greedy decoding was used (see section 5.1.1.1). However, greedy decoding can only produce one sample. Therefore, for generating a set of samples we resort to the use of sampling, which produces  $k$  inde-

---

**Algorithm 1** The REINFORCE algorithm with multiple-samples per data point.

---

**Require:**

A pre-trained policy ( $\pi_\theta$ ).

**Input:** Input ( $X$ ), ground-truth expressions ( $Y$ ),

**Output:** A fine-tuned policy with REINFORCE with multiple-samples.

**Training Steps:**

**while** not converged **do**

    Produce a mini-batch of size  $N$  from  $X$  and  $Y$ .

**for** each element in  $N$  **do**

        Generate  $K$  full sequences of actions:

$\{\hat{y}_1, \dots, \hat{y}_T \sim (\hat{y}_1^{RS}, \dots, \hat{y}_T^{RS})\}_1^K$ .

        Observe the sequence rewards and calculate the baseline  $b_i = \frac{1}{K-1} \sum_{i \neq j} r(\hat{y}^j)$ .

**end for**

    Calculate the loss according to Eq. (4.8).

    Update the parameters of network  $\theta \leftarrow \theta + \alpha \nabla_\theta \mathcal{L}_\theta$ .

**end while**

---

pendent samples by sampling from the model’s distribution. Algorithm 1 summarizes the required steps for the proposed approach.

### 4.2.3 Reward Configuration

The standard training for REG poses two uncommon challenges for RL. First, the action space in REG problems is a high-dimensional discrete space that it is intractable, while in the classic RL paradigm the common scenario is a smaller discrete action space (e.g. games (71)), or a relatively low dimension continuous space of actions (e.g. robotics (57)). Hence, the first important factor is the *search strategy* for generating the sequence of actions.

For generating sequences and set of sequences two sampling strategies were used. The first is *beam search*, that finds the most likely sequence by performing a greedy breadth-first search over a limited search space. Specifically, each candidate sequence is expanded from left to right selecting all possible tokens from the vocabulary at a time. From this set, the *top-k* candidate sequences with the highest probabilities are selected, and the beam search process continues until the *top-k* candidates with the highest probability are returned. The second strategy is random sampling, which randomly samples from the model’s distribution at every time-step until the end of the sequence token is produced.

Balancing between *exploration* and *exploitation* is a major challenge in RL. For

instance, it may be required for an agent to pick an action associated with the highest expected reward (i.e. exploitation). However, in this scenario it may fail to learn more rewarding actions. Therefore exploration, that is the choice of new actions and the visit of new states, may also be beneficial. Beam search focuses on producing high probability sequences and therefore is considered as an exploitation strategy, while random sampling introduces more diverse sequences and thus contributes towards the exploration of the action states. However, due to the fact that the actions are being sampled from the model being optimized the exploration is de facto limited.

### 4.3 Minimum Risk Training for Referring Expression Generation

Similar to the proposed RL object, minimum risk training is an objective that is computed over a set of sequences. Specifically, MRT minimizes the value of a given task-specific cost function, i.e. risk, over the training data at sequence level. Let  $x$  denote a fixed-size representation of the input, then the set  $\mathcal{Y}(\mathbf{x}^{(s)})$  denotes the set of all possible referring expressions generated by the model with parameters  $\theta$ . For a given candidate sequence  $\mathbf{y}'$  and ground truth referring expression  $\mathbf{y}$ , MRT defines a cost function  $\Delta(\mathbf{y}', \mathbf{y})$  which is the semantic distance between  $\mathbf{y}'$  and the standard  $\mathbf{y}$ . The cost function can be any function that captures the discrepancy between the model's prediction and the ground truth. Formally, the objective function of MRT is the following:

$$(4.9) \quad \mathcal{L}_{\text{MRT}} = \sum_{n=1}^N \mathbb{E}_{\mathcal{Y}(\mathbf{x})} \Delta(\mathbf{y}', \mathbf{y}^{(n)}).$$

where  $\mathbb{E}_{\mathcal{Y}(\mathbf{x})}$  denotes the expectation over the set of all possible candidate sequences  $\mathcal{Y}(\mathbf{x}^{(n)})$ . However, as previously mentioned enumerating and scoring candidate sequences over the entire space is intractable. Instead, we sample a subset  $\mathcal{S}(\mathbf{x}) \subset \mathcal{Y}(\mathbf{x})$  to approximate the probability distribution, and formalize the objective function as:

$$(4.10) \quad \mathcal{L}_{\text{MRT}} = \sum_{s=1}^S \sum_{\mathbf{y}' \in \mathcal{S}(\mathbf{x}^{(s)})} \frac{p(\mathbf{y}' | \mathbf{x}^{(s)})}{\sum_{\mathbf{y}^* \in \mathcal{S}(\mathbf{x}^{(s)})} p(\mathbf{y}^* | \mathbf{x}^{(s)})} \Delta(\mathbf{y}', \mathbf{y}^{(s)})$$

The MRT objective minimizes the expected value of a cost function which enables the optimization of REG models with respect to specific evaluation metrics of the task. In this chapter the use of various REG evaluation metrics such as CIDER and BLEU and a combination of those is explored (e.g. CIDER + BLEU). Furthermore, for the construction of the subset of the candidate sequences both online and offline generation were considered. However, empirically it was found that offline generation leads to inferior performance and thus was not included. For the generation of sets of expressions, random sampling and beam search as search strategies (see Section 4.2.3) were considered.

## 4.4 Combined objectives

The motivation of the loss combination is to maintain good token-level accuracy while optimizing on the sequence-level. In other words, using an evaluation metric as a reward can suppress the probability of the words that do not increase the metric score, and thus concentrate the distribution to a single point. Thus, a combined objective is explored in order to scale the peakiness of the output distribution. Specifically, the weighted combination of MLE with RL objective is defined as follows:

$$(4.11) \quad L_{weighted_{RL}} = (1 - \alpha) * L_{mle} + \alpha * \hat{L}_{rl},$$

Equivalently, combining the MRT objective (Equation 4.10) with MLE we have:

$$(4.12) \quad L_{weighted_{MRT}} = (1 - \alpha) * L_{mle} + \alpha * \hat{L}_{MRT},$$

where  $\alpha$  is a scaling factor controlling the difference in magnitude between the combined objectives.



## 4.5 Experimental design and results

### 4.5.1 Implementation Details

The visual features, datasets and the implementation details of the language models are described in Section 3.2.1. Both the RL and MRT models are trained according to the following scheme: first the language model is pretrained using MLE, optimized with Adam (45). At each epoch, the model is evaluated on the validation set and the model with the best CIDEr score is selected as an initialization for RL and MRT training. Then the model is optimized with either RL or MRT.

### 4.5.2 Evaluation

The quality of the output is measured by standard automatic metrics that have been used in REG (67, 116, 119) and they are described in Section 3.2.2. Furthermore, in order to measure the diversity the following metrics are reported: (1) the average length of referring expressions (ASL); (2) the number of unique words of the generated corpus; (Voc) and (3) the average number of unique bigrams per 1000 bigrams (TTR) (68).

### 4.5.3 Evaluating different RL training configurations

First, a number context-dependent normalization factors that affect the RL training are explored. Regarding the reward configuration (see Section 4.2.3) the following factors are explored: (1) which reward function to use to evaluate the sequences; and (2) which search strategy will be used to sample the actions from the policy.

**Reward Function:** First various evaluation measures are compared as reward functions, namely CIDEr, BLEU and METEOR as well as metrics combinations. A summary of the results is given in Table 4.1, where RL stands for the REINFORCE algorithm. The performance of the MLE model that used for the initialization of the RL training is also presented. As expected, optimizing towards a particular evaluation metric during training leads to an increase on that particular metric during testing. However, the

RefCOCO						
	testA			testB		
Training Metric	CIDEr	$BLEU_1$	RL + METEOR	CIDEr	$BLEU_1$	METEOR
MLE	0.762	0.490	0.177	1.332	0.523	0.208
RL + CIDEr	<b>0.978</b>	<b>0.556</b>	<b>0.211</b>	<b>1.498</b>	<b>0.536</b>	<b>0.229</b>
$BLEU_1$	0.811	0.512	0.190	1.342	0.501	0.211
RL + METEOR	0.762	0.489	0.178	1.331	0.522	0.209
RL + CIDEr+ $BLEU_1$	0.914	0.534	0.202	1.422	0.527	0.223
RefCOCO+						
	testA+			testB+		
Training Metric	RL+CIDEr	$BLEU_1$	METEOR	CIDEr	$BLEU_1$	METEOR
MLE	0.633	0.444	0.167	0.710	<b>0.373</b>	0.159
RL + CIDEr	<b>0.847</b>	0.500	0.203	<b>0.980</b>	0.288	0.169
RL+ $BLEU_1$	0.760	0.480	0.189	0.914	0.299	0.163
RL + METEOR	0.729	0.442	0.179	0.932	0.321	0.169
RL + CIDEr+ $BLEU_1$	0.845	<b>0.517</b>	<b>0.207</b>	0.979	0.299	<b>0.171</b>

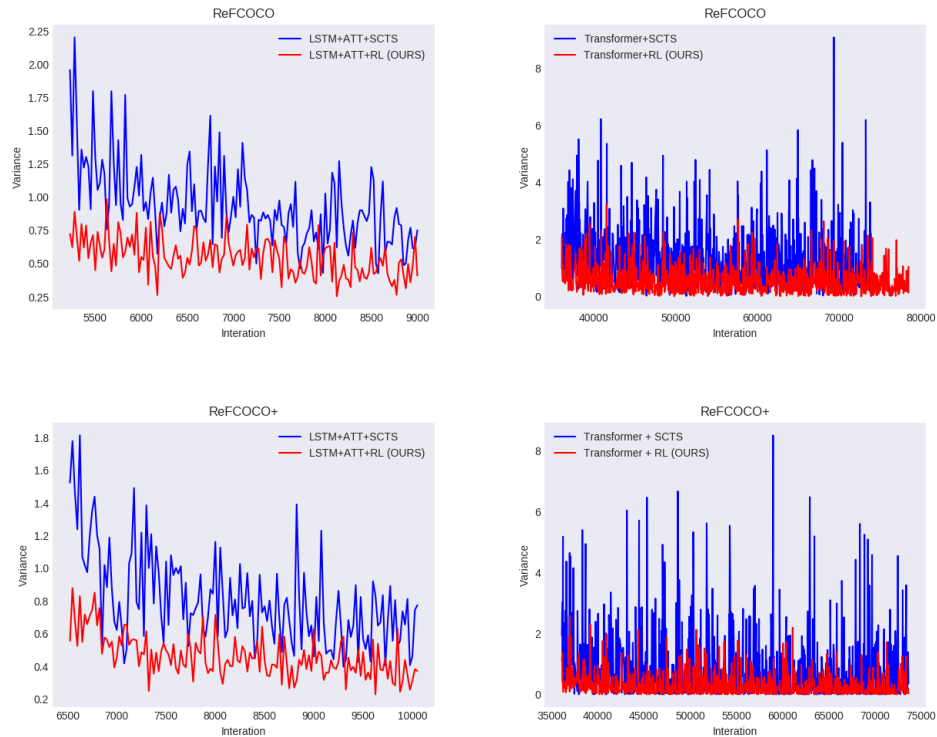
**Table 4.1:** Performance of different reward functions for the LSTM model. When the language model is optimized with the CIDEr metric, a significant increase to all other evaluation metrics is observed. All models were decoded using greedy decoding. The performance of the seed model is also reported. The best overall values for each metric are emphasized with bold.

benefits are not comparable with those gained when optimizing CIDEr. Specifically, CIDEr optimization leads to improvements in scores for all other metrics as opposed to directly optimize them. A notable exception is the combination of CIDEr+BLEU where BLEU score is higher compared to optimizing only for CIDEr. Therefore, for the rest of the this, all RL models are based on CIDEr optimization.

Method	testA			testB			testA+			testB+		
	BLEU	METEOR	CIDEr	BLEU	METEOR	CIDEr	BLEU	METEOR	CIDEr	BLEU	METEOR	CIDEr
MLE	0.542	0.200	0.841	0.614	0.258	1.507	<b>0.481</b>	0.179	0.715	<b>0.409</b>	<b>0.173</b>	0.829
RL+ RS	0.569	0.222	0.954	0.625	0.277	1.564	0.469	0.185	0.745	0.286	0.163	0.913
RL + BS	0.561	0.217	0.946	0.617	0.270	1.549	0.465	0.184	0.743	0.277	0.160	0.901
SCTS+RS	<b>0.593</b>	<b>0.231</b>	<b>1.012</b>	<b>0.638</b>	<b>0.290</b>	<b>1.607</b>	<b>0.481</b>	<b>0.194</b>	<b>0.809</b>	0.282	0.165	<b>0.942</b>
SCTS+GD	0.583	0.227	0.995	0.635	0.279	1.585	0.461	0.185	0.761	0.276	0.163	0.934

**Table 4.2:** Results of different search strategies for reward computation and variance reduction for the LSTM model. “RS” stands for random sampling, while “BS” refers to beam search and “GD” for greedy decoding. “SCTS” refers to self-critical training. Shaping denotes that we used reward shaping.

**Action sampling strategy:** So far the words were sampled using random sampling. Next, beam search and random sampling are compared as sampling strategies. The



**Figure 4.2:** Gradient variance of the proposed RL objective compared to the SCST for the proposed attention and transformer model.

results are shown in Table 4.2. Although beam search (with width of 2) has been the de facto decoding strategy for neural REG systems, it produces inferior results when compared to random sampling. Due to the deterministic nature of beam search, the sampled sequences are often duplicates and thus uninformative for the gradient estimation, while the stochasticity of sampling generates sequences with exploratory usefulness for the gradient estimation and it results in more diverse expressions.

#### 4.5.4 Results of the proposed RL objective

This section evaluates whether the proposed RL objective reduces the variance of the gradient compared to self-critical sequence training. High variance increases sample complexity and can impede effective learning, hence the reduction of the variance introduces better learned models (65). Although both techniques lead to unbiased estimators of the gradient, the proposed method results in lower gradient variance for both language models that were tested. Interestingly, SCST has much higher gradient variance than the proposed RL objective during the first epoch of training. It is hypo-

Model	RefCOCO				RefCOCO+			
	testA		testB		testA		testB	
	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$	CIDEr	$BLEU_1$
LSTM +ATT + MLE	1.033	0.594	1.552	0.609	0.884	0.512	0.858	0.424
LSTM + ATT + SCST	1.089	0.597	1.565	0.570	1.065	<b>0.563</b>	1.054	0.323
LSTM + ATT+ RL(OURS)	1.204	0.636	1.646	0.605	<b>1.077</b>	<b>0.563</b>	1.074	0.323
LSTM + ATT+ MRT	1.054	0.612	1.570	0.615	<b>0.952</b>	<b>0.569</b>	0.987	0.373
Transformer	0.938	0.529	1.464	0.586	0.938	0.529	0.913	<b>0.424</b>
Transformer + SCST	1.255	0.650	1.710	0.650	0.967	0.532	0.974	0.308
Transformer + RL(OURS)	<b>1.261</b>	<b>0.665</b>	<b>1.732</b>	<b>0.656</b>	1.020	0.546	1.003	0.294
Transformer + MRT	<b>1.071</b>	<b>0.608</b>	<b>1.591</b>	<b>0.610</b>	0.965	0.534	0.945	0.395

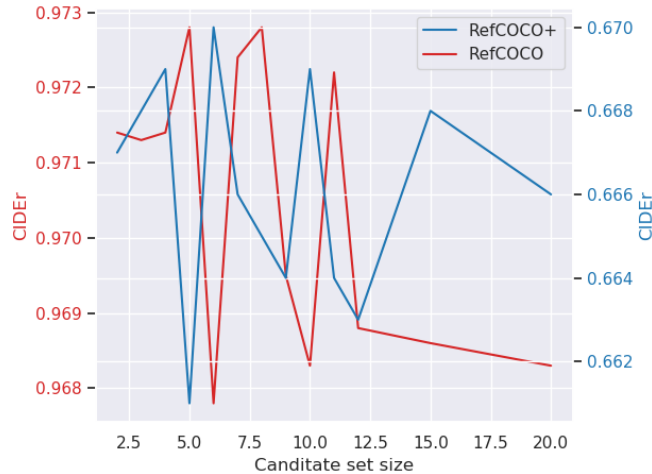
**Table 4.3:** Performance of the best attention (denoted as LSTM +ATT) and transformer model (denoted as Transformer ) trained with maximum likelihood estimation (denoted as MLE), self-critical sequence training (denoted as SCST) and the proposed RL objective (denoted as RL (OURS)). The models were greedily decoded.

thesized that the samples drawn from the model’s distribution score lower than the sentences produced by greedy decoding.

Table 4.3 presents the results on RefCOCO and RefCOCO+ for the proposed attention model and the transformer model optimized with the proposed RL training strategy and SCST. Both RL based models are fine-tuned from the same pre-trained model. Again, in order to determine whether the difference in scores between the two RL methods was statistically significant, a two-tailed t-test was performed. Specifically, when a model is optimized with the proposed RL objective achieves higher CIDEr scores than SCST. The score difference was statistically significant. Second, the  $BLEU_1$  score difference was statistically significant in RefCOCO and RefCOCO+ testB. Third, it is observed that the attention model achieves higher scores when trained with MLE in both datasets compared to transformer. However, when both models are trained with RL, the transformer presents higher scores than the attention LSTM in RefCOCO.

#### 4.5.5 Evaluating Minimum Risk Training for REG

Training with MRT requires generating and scoring multiple candidate referring expressions for each input. Thus, two factors are explored: (1) which search strategy should be used to generate the candidate sequences; (2) and how many sequences should be generated for one input. It was found that random sampling performs better than



**Figure 4.3:** Validation set CIDEr scores for different candidate set sizes for the MRT model. Best viewed in color.

beam search both in terms of CIDEr scores and is considerably faster. Thus, Figure 4.3 compares different set sizes on the validation set when random sampling is used. For RefCOCO a set size of 5 was chosen, while for RefCOCO+ 8. Table 4.3 presents the results on the test set. When optimizing the REG model with MRT consistent improvements in terms of CIDEr and  $BLEU_1$  scores are observed across datasets compared to the MLE trained model. The difference in scores between the two objectives was found statistically significant.

#### 4.5.6 Comparison between objectives

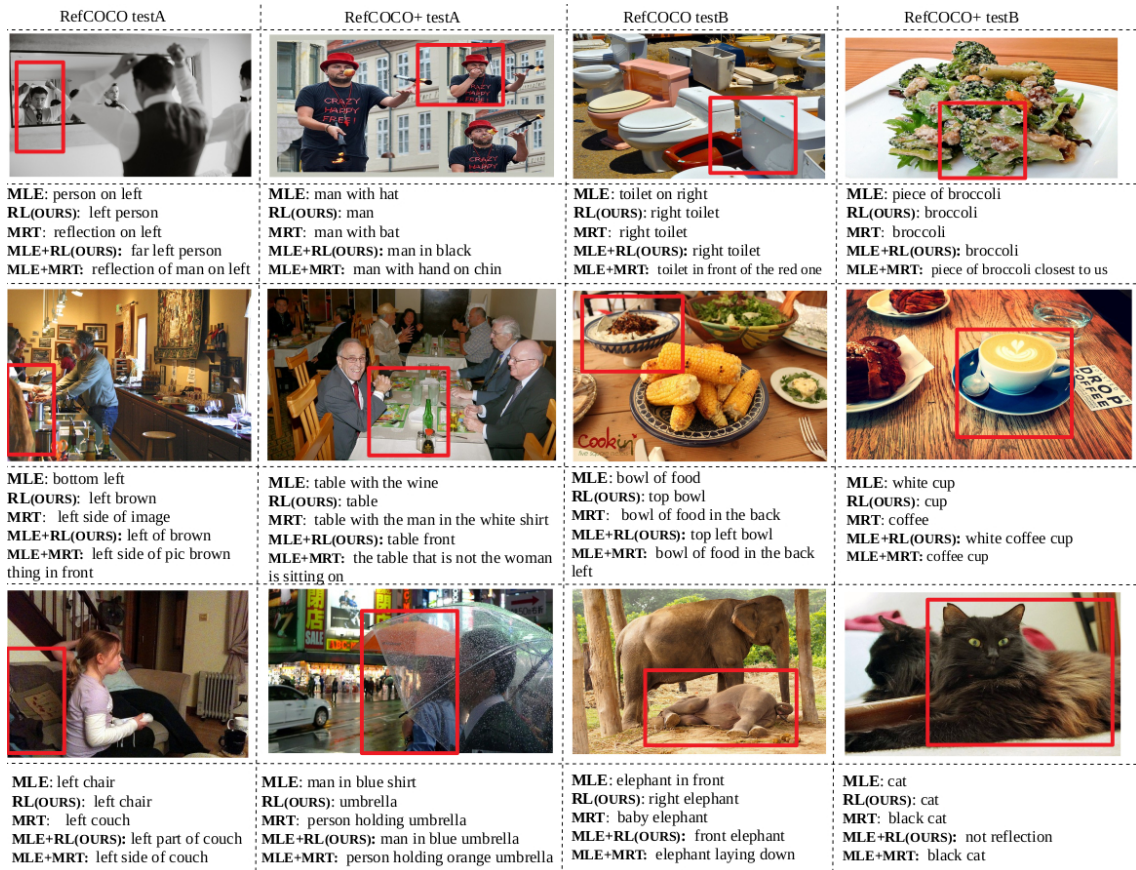
Tables 4.4, 4.5 show all sequence level optimization methods used. When analyzing the effect that different training methods have on diversity of the referring expressions a few clear patterns can be observed: (1) the proposed RL objective has the lowest diversity and highest repetition among all models (i.e. TTR scores); (2) out of the 4 different test sets, the proposed objective has the highest CIDEr scores when compared to MLE and MRT training; (3) combining the proposed RL objective with MLE improves slightly the accuracy, naturalness and diversity of the produced referring expressions. Still, however, the diversity is considerably lower than MLE and MRT. (4) Minimum risk training improves over MLE in all tests sets. (5) MRT has the highest diversity compared to the other training strategies; (6) combining the MRT loss with MLE further improves

	RefCOCO testA				
	CIDEr	BLEU	Avg. Length	Unique words	TTR
MLE	0.938	0.529	<b>2.52</b>	246	0.245
RL(ours)	1.261	0.665	1.40	202	0.173
MRT	1.070	0.608	2.66	243	0.236
MLE+ RL (ours)	<b>1.275</b>	<b>0.669</b>	2.41	227	0.188
MLE+MRT (ours)	1.193	0.642	<b>2.92</b>	<b>275</b>	<b>0.290</b>
	RefCOCO testB				
MLE	1.464	0.586	2.76	275	0.259
RL(ours)	1.732	0.656	2.41	209	0.201
MRT	1.591	0.610	2.52	277	0.233
MLE + RL (OURS)	<b>1.825</b>	0.658	2.68	230	0.226
MLE + MRT	1.739	<b>0.662</b>	<b>2.80</b>	<b>306</b>	<b>0.286</b>

**Table 4.4:** System results for RefCOCO : CIDEr and BLEU scores; average sentence length (ASL); vocabulary size (Voc); mean-segmented bigram ratio (TTR); RL (ours) denotes the proposed RL objective; MRT denotes minimum risk training.

	RefCOCO+ testA				
	CIDEr	BLEU	Avg. Length	Unique words	TTR
MLE	0.938	0.529	2.80	357	0.337
RL(ours)	1.020	0.546	1.92	242	0.193
MRT	0.965	0.534	2.66	324	0.305
MLE+ RL (ours)	<b>1.035</b>	0.548	2.54	306	0.275
MLE+MRT (ours)	0.986	<b>0.563</b>	<b>3.04</b>	<b>368</b>	<b>0.340</b>
	RefCOCO+ testB				
MLE	0.913	0.424	2.92	428	0.377
RL(ours)	1.003	0.294	1.58	277	0.243
MRT	0.945	0.395	2.86	387	0.340
MLE + RL (OURS)	<b>1.036</b>	0.323	2.55	368	0.336
MLE + MRT	0.952	<b>0.419</b>	<b>3.11</b>	<b>431</b>	<b>0.396</b>

**Table 4.5:** System results for RefCOCO+ : CIDEr and BLEU scores; average sentence length (ASL); vocabulary size (Voc); mean-segmented bigram ratio (TTR); RL (ours) denotes the proposed RL objective; MRT denotes minimum risk training.



**Figure 4.4:** Examples of objects and expressions drawn from both RefCOCO and RefCOCO+ datasets. The target object is highlighted with a red box.

the diversity and naturalness of the generated referring expressions.

Examples of generated REs are illustrated in Figure 4.4. In all images presented in Figure 4.4, we observe that the proposed MLE + MRT model improves over all compared training objectives in inferring more pragmatically adequate referring expressions by using, for example, precise appearance and location attributes (e.g. “man with hand on chin” and “left side of pic brown thing in front”).

## 4.6 Conclusions

In this chapter the problem of optimizing REG models with sequence level objectives was investigated. There are two reasons for which it is desired the models to be trained on sequence level: (1) to avoid the loss-evaluation metric mismatch coined as *exposure bias*, that is during training the model utilizes a word-level loss, while during generation its goal is to generate an expression that improves sequence-level metrics; and (2) to

expose the model to its own predictions during training, as opposed to let the model be exposed to its own prediction only during testing. Reinforcement learning have been proposed to solve the aforementioned problems. However, there two major limitations in applying RL techniques in REG: (1) the gradient exhibits high variance; (2) applying RL techniques is not a trivial task since without proper context-dependent normalization the training is unstable.

In order to reduce the variance of the gradient a variation of the popular REINFORCE algorithm that utilizes multiple samples per input to normalize the reward was proposed. It was showed that the proposed approach reduces the variance of the gradient more effectively compared to the self-critical sequence training. Furtermore, in order to evaluate the effectiveness of the proposed approaches human evaluation was conducted. Specifically, all human evaluation experiments were conducted at Amazon Mechanical Turk. An Amazon Mechanical Turk worker was able to participate in this human evaluation experiment only if the worker was a native speaker of the English language, was located in English- speaking countries, had an approval rate of 99% and had successfully completed 1000 tasks. Each worker was only allowed to participate in a task only once. In each task the workers were presented with an image and an expression and were asked to draw a box around the object that they believe is best described by that expression. If two workers chose the correct object, then the expression was considered successful. For each test set 60 objects were randomly selected. For each test set 180 unique ratings were collected. The human evaluation results on RefCOCO and RefCOCO+ dataset establish a new state-of-the art. The proposed approach improve the results in ReFCOCO testA and testB from 76.95% to 81.66% and from 78.10% to 83.33% respectively. While in RefCOCO+ testA it improves the best results from 58.85% to 83.33%. Furthermore, a comprehensive comparison of different aspects of configuring REG models with RL training was presented. It was found that (1) that directly optimizing the CIDEr metric is highly effective; (2) random sampling is a better search strategy than beam search to sample actions; and (3) using random sampling with self-critical training improves CIDEr scores. Another major limitation that stems from REG models trained with RL is that the generated referring expressions lack in diversity

---



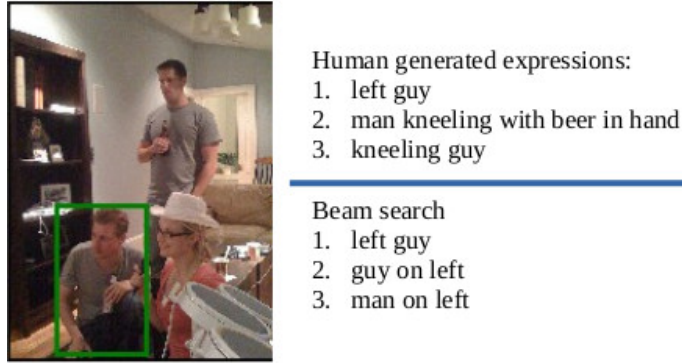
and naturalness due to the deficiencies in the generated words distribution, smaller vocabulary size and the receptiveness of the frequent words. In order to address this issue two solutions were proposed. First, to combine RL objectives with MLE. It was showed that combining the two objectives is beneficial to the training, resulting in higher CIDEr scores and diversity. However, there is a considerable gap between MLE and RL methods w.r.t. to diversity. Thus, as an alternative method to optimize REG model in sequence level, a novel objective based on minimum risk training was proposed. It was found that that MRT produces superior results in terms of diversity of the referring expressions compared to RL training. Furthermore, it was showed that MRT combined with MLE produces superior results in terms of naturalness and diversity of the referring expressions compared to MLE.

## *Decoding strategies*

---

Existing efforts in neural REG focus on training objectives that promote resemblance to the ground truth sentences in order to reduce ambiguity. Secondly, due to their autoregressive nature, exact inference for generating the most likely output is intractable. Thus, it is necessary to resort to approximate search algorithms such as beam search (46). However, despite the widespread adaptation of beam search, it has been found that the output decoded with beam search lacks in diversity (37, 101, 104). As shown in Figure 5.1, beam search produces near identical expressions, with minor morphological variations (37, 100). Furthermore, although multiple referring expressions are potentially correct for one target object, existing efforts produce a single referring expressions.

Therefore, this chapter studies the generation of sets of referring expressions. Specifically, this chapter investigates the effect of different decoding strategies by comparing their performance along the entire quality-diversity space. The importance that NLG systems place on these two criteria, is application dependent. For example, the goal of an open domain dialogue generation system is to be able to converse for a variety of topics and thus places more weight in the diversity of the output (51). However, in REG the most important attribute of the output is to successfully identify the target object. Thus, generating a set of expressions is useful only if it does not come on the expense of the quality. Therefore, this chapter presents the first large-scale human evaluation to measure how the hyperparameters of each decoding algorithm, affect the diversity and the quality of sets of referring expressions.



**Figure 5.1:** An example image associated with the top three referring expressions decoded with standard beam search and those provided by humans annotators. The target object is highlighted with the green box.

## 5.1 Contributions

Therefore the contributions of this chapter are the following:

- Extends the investigation to the generation of sets of referring expressions in order to reproduce the diversity found in human written referring expressions. In particular, the first detailed comparison of how the hyperparameters of commonly-used decoding strategies affect the quality-diversity trade-off is presented. This contribution is described by [Panagiaris, Hart and Gkatzia \[73\]](#).
- Presents the first large-scale human evaluation that measures the impact that diversity has on the quality of sets of referring expressions. This contribution is described by [73](#).

### 5.1.1 Maximization-based decoding methods

#### 5.1.1.1 Greedy Decoding

Greedy decoding (GD) can be seen as a naive inference method for conditional language models ([121](#)). It chooses the most likely token of the sequence, in a left to right manner, under the conditional probability:

$$\hat{x}_t = \operatorname{argmax}_{x_t} P(x_t | x_{<t}, I, r)$$

The process continues until the end symbol is produced. Although it is computationally efficient, it can often lead to sub-optimal solutions (10). A significant drawback of this approach is that, the high-probability choices in earlier generation steps, can lead to an overall low likelihood sequence due to low probabilities choices later on.

### 5.1.1.2 Beam Search

Beam search (BS) is an inference algorithm that explores in a greedy left-right manner the search space (121). Instead of extending a single hypothesis, at each time step, it extends a set of  $K$  hypotheses  $H_t$ :

$$(5.1) \quad \mathcal{H}_t = \{(x_1^1, \dots, x_t^1), \dots, (x_1^K, \dots, x_t^K)\}.$$

The next set of partial hypotheses is created by expanding all the hypotheses in  $\mathcal{H}_t$  with each token from the vocabulary  $V$ . Then, each candidate hypothesis  $h_{x_t^i}^i$  from  $H_t$  is scored as:

$$(5.2) \quad s(\tilde{h}_{v_j}^i) = s(h_{\tilde{y}_t^i}^i) + \log p(v_j | \tilde{x}_{\leq t}^i).$$

The  $K$  highest ranked hypotheses are selected as the new candidate set to be expanded in the next step. Among the top hypotheses, those whose the last token is the special EOS token are no longer expanding. The remaining hypotheses continue to expand, however, with  $K$  reduced by the number of complete hypotheses. This process terminates until  $K$  reaches zero, and the best completed hypotheses are returned.

The space that beam search performs is the union of all the current hypotheses in  $\mathcal{H}^k$ . Thus, the  $K$  decoded sequences are from the same high-likelihood subspace. Consequently, generating a set of notable different expressions for a target object is not trivial.

### 5.1.1.3 Diverse Beam search

Diverse Beam Search (DBS) (100) is a variant of beam search that tries to alleviate the redundancy of the search lists. DBS introduces a dissimilarity term  $\theta$  that measures the difference between the current hypotheses with those produced in the previous step. It

achieves that by augmenting the log-likelihood before re-ranking. More formally, each candidate hypothesis  $\tilde{h}_{\leq t}^i$  is scored as:

$$s(\tilde{h}_{\leq t}^i) = s(h_{\leq t}^i) + \lambda\theta(h_{\leq t}^i, H_{t-1}).$$

where  $\lambda$  is a factor that regulates the strength of diversity. Another important hyperparameter is the dissimilarity function  $\theta$ . We follow Vijayakumar et al. [101] and as dissimilarity function we use the Hamming distance that was reported to perform best.

A limitation that stems from this approach is that the fixed diversity strength is not optimal in every scenario. 100 reported that complex images benefit more from diversity-promoting inference than simpler images.

### 5.1.2 Sampling-based decoding methods

An alternative to decoding based on maximization is the introduction of some element of randomness by sampling from the model’s learned distribution. In this scenario, at each time step  $t$  the next word is randomly drawn from the conditional language model as:

$$(5.3) \quad x_i \sim P(x|x_{1:i-1}, I_i, r_i)$$

While output generated using this process avoids repetitions, it can become incoherent by sampling from model’s low confidence zones (37). REG is a low tolerance task; only one word is enough for an unsuccessful referring expression (e.g. color or location words). To avoid sampling words from the tail of the distribution, which contains a large number of tokens assigned with low probability, three solutions have been proposed: (1) sampling with temperature; (2) top- $k$  sampling; (3) and nucleus sampling.

#### 5.1.2.1 Sampling with temperature

One common approach to control the entropy of the distribution is the use of temperature (21, 23):

$$(5.4) \quad p(x = V_l|x_{1:i-1}, I, r) = \frac{\exp(u_l/T)}{\sum_{l'} \exp(u_{l'}/T)}.$$

The use of temperature  $T \in [0, 1)$  reduces the risk of sampling words with very low probability, by skewing it towards high-probability zones (37).

### 5.1.2.2 Top- $k$ Sampling

Top- $k$  sampling that was proposed by [Fan, Lewis and Dauphin \[19\]](#), is an intuitive solution that truncates the distribution by maintaining a subset of high-probability tokens. At each time step a fixed number of  $k$  words are selected that maximize  $p' = \sum_{x \in V^{(k)}} P(x|x_{1:i-1}, I, r)$ . Then, the next words are drawn from the top- $k$  vocabulary  $V^{(k)} \subset V$  based on their relative probabilities. Formally, the next words are drawn as follows:

$$(5.5) \quad P^*(x|x_{1:i-1}, I, r) = \begin{cases} P(x|x_{1:i-1}, I, r) / p' & \text{if } x \in V^{(k)} \\ 0 & \text{otherwise} \end{cases}$$

### 5.1.2.3 Nucleus Sampling

An alternative to top- $k$  sampling is nucleus sampling proposed by [Holtzman et al. \[37\]](#). The fundamental difference between those two sampling strategies is that nucleus sampling instead of having a fixed number of tokens as subspace, it uses those tokens whose cumulative probability mass surpass a pre-define threshold  $q$ . Thus, the next words are drawn from the vocabulary  $V^{(q)} \subset V$  which is the smallest set that:

$$(5.6) \quad \sum_{x \in V^{(q)}} P(x|x_{1:i-1}, I, r) \geq q.$$

sectionGeneration of a set of REs

Decoding Method	Hyperparameter
Random Sampling (RS)	Temperature $T$
top- $k$ Sampling	he number $k$ of tokens to be kept.
Nucleus Sampling (NS)	The probability threshold $q$
Diverse beam search (DBS)	The diversity strength parameter $\lambda$

Beam search (BS)	Temperature $T$
------------------	-----------------

**Table 5.1:** The hyperparameter that controls the quality-diversity trade-off for each of the decoding strategies used in this work.

## 5.2 Experimental design and results

### 5.2.1 Implementation Details

The experiments in this section explore how decoding algorithms affect the accuracy-diversity trade-off. The language model that was chosen is the transformer model (see Section 3.1.3) that achieved state-of-the-art results in human evaluation for the one-shot generation. The visual features, datasets and the implementation details of the transformer are described in Section 3.2.1. As the focus of this chapter is the how the decoding parameters affect the generation of sets of referring expressions the cross-entropy loss (see Section 3.1.4) was used. Table 5.1 shows the diversity parameter that controls the accuracy-diversity trade-off for each of the employed decoding strategies.

Secondly, most of the published models are trained to generate a single referring expression, thus the decoding strategies were adapted to generate a set of REs. In particular, two approaches were used to generate a set: (1) for the randomization-based algorithms (e.g. random sampling), a set of referring expressions is constructed by randomly sampling from the model’s learned distribution; and (2) for normal and diverse beam search the beam width was used to generate the set.

### 5.2.2 Evaluation

In order to evaluate a *set* of referring expressions two criteria are required to be taken into consideration: accuracy and diversity. For the former, the commonly used approach is to average a similarity score (41), e.g. CIDEr, over the set. Evaluating the accuracy of a particular system is not sufficient to reflect the overall performance of

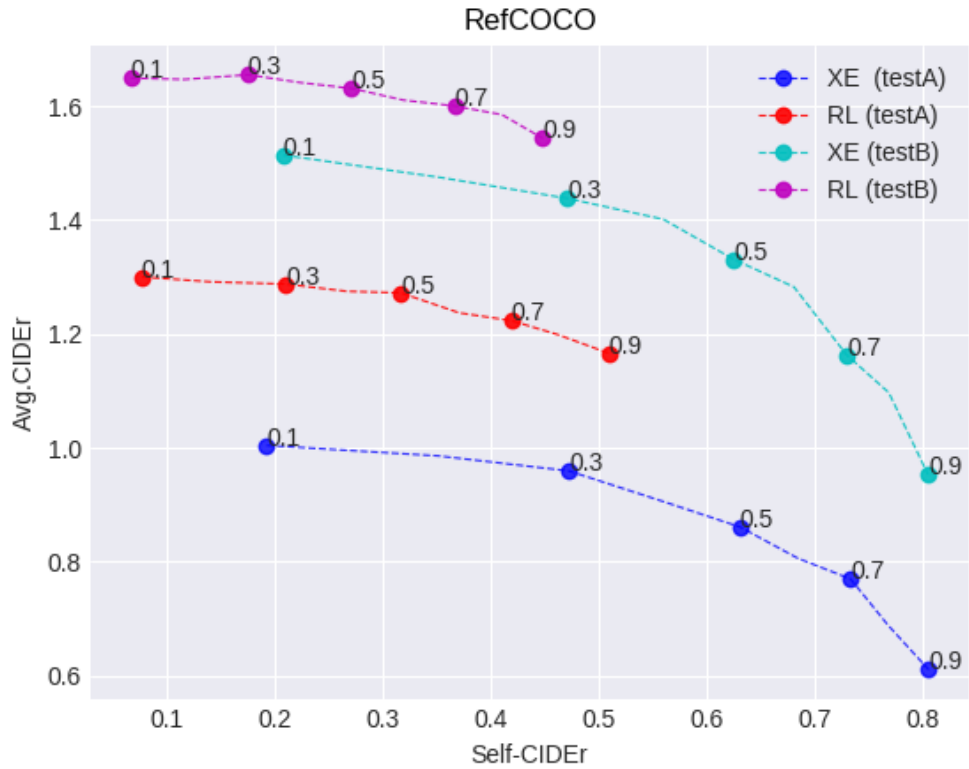
a model; the diversity of the output should also be considered. The diversity of a set is computed using Self-CIDEr (104), that computes a diversity score by calculating the eigenvalues of a kernel matrix that contains similarities scores (i.e. CIDEr) for all sentences pairs within the set (104).

Furthermore, human evaluation was conducted on Amazon Machine Turk and human judges were asked to rate the diversity and the quality of a set of referring expressions. The human evaluation experiments were conducted at Amazon Mechanical Turk. An Amazon Mechanical Turk worker was able to participate in this human evaluation experiment only if the worker was a native speaker of the English language, was located in English-speaking countries, had an approval rate of 99% and had successfully completed 1000 tasks. Each worker was only allowed to participate in a task only once. In each task the workers were presented with an image and a set of expressions. For each expression in the set, three workers were asked whether or not the expression describes the object unambiguously. A referring expression was considered successful if two workers found that the expression unambiguously describes the object. Then the workers were required to rate the diversity of the set on a 5-point Likert scale, where 1 indicates that the expressions are identical, and 5 that the expressions are significantly different with one another. In the instruction given to the workers, diversity was referred to different words, phrases, sentence structures, semantics or other factors that impact diversity. The diversity score for each set is the average score given by the 3 workers. In total 25 target objects were randomly selected and 5 expressions were generated for each object and ratings from 1275 workers were elicited.

### 5.2.3 Results for Random Sampling-based Decoding Methods

The aim of the experiments in this sections is to investigate how the temperature affects the accuracy-diversity trade-off for random sampling. As illustrated in Figures 5.2, 5.3 higher sampling temperatures result in both an increase in Self-CIDEr scores and a reduction in average CIDEr scores. Interestingly, using CIDEr reward to fine-tune the model will drastically reduce Self-CIDEr, while will increase the average CIDEr score.

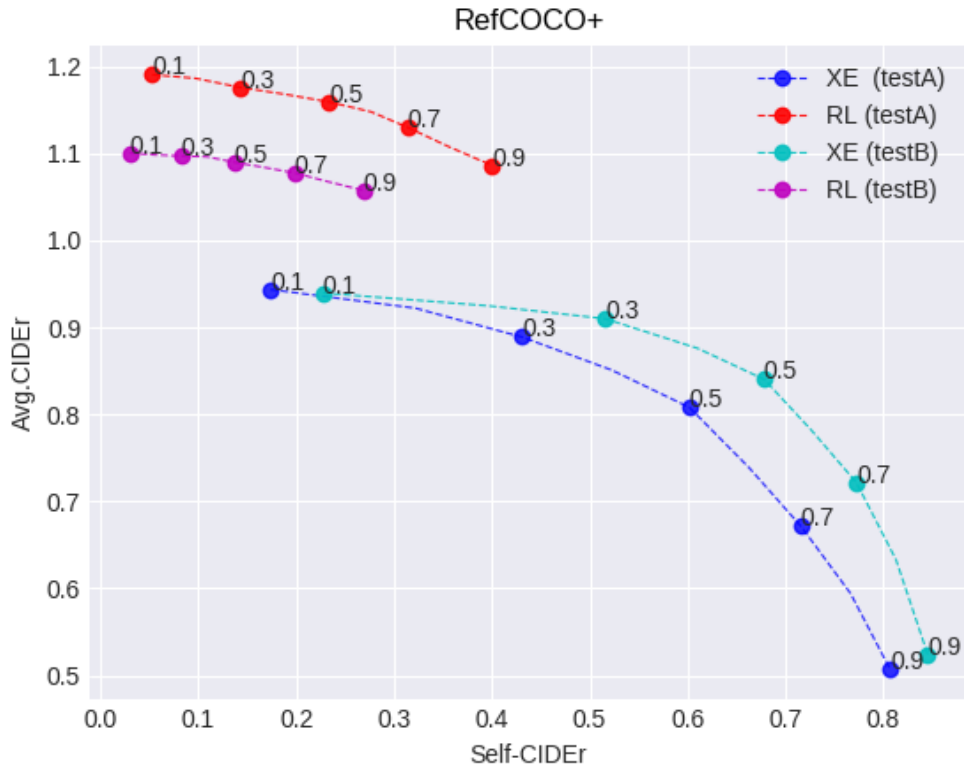




**Figure 5.2:** Self-CIDEr and average CIDEr scores for random sampling with different temperature values for RefCOCO dataset. The language model used is the proposed transformer trained with cross-entropy (XE) and fine-tuned with the proposed RL objective.

Optimising the CIDEr reward encourages syntactic similarity between the generated expressions and the ground truth expressions which leads to low diversity. Hence, granted that the reinforcement learning objective drastically affects the diversity of the produced expressions, it will no longer be used in the rest of the experiments of this chapter. To illustrate the differences between the two objectives and how different temperature values affect the output, example objects with the corresponding expressions generated by each model are presented in Figure 5.6. First, it is observed that for both objectives when the sampling temperature is set to 0.1, the text is highly repetitive, mimicking greedy search. Furthermore, when the temperature is set to 0.9, the model trained with cross entropy produces output that is less fluent and incoherent (see Figure 5.6).

Top- $k$  and nucleus sampling have become an alternative to random sampling. Both strategies sample from a truncated distribution. The difference between the two is how they truncate the distribution; top- $k$  sampling keeps a fixed number of  $k$  tokens

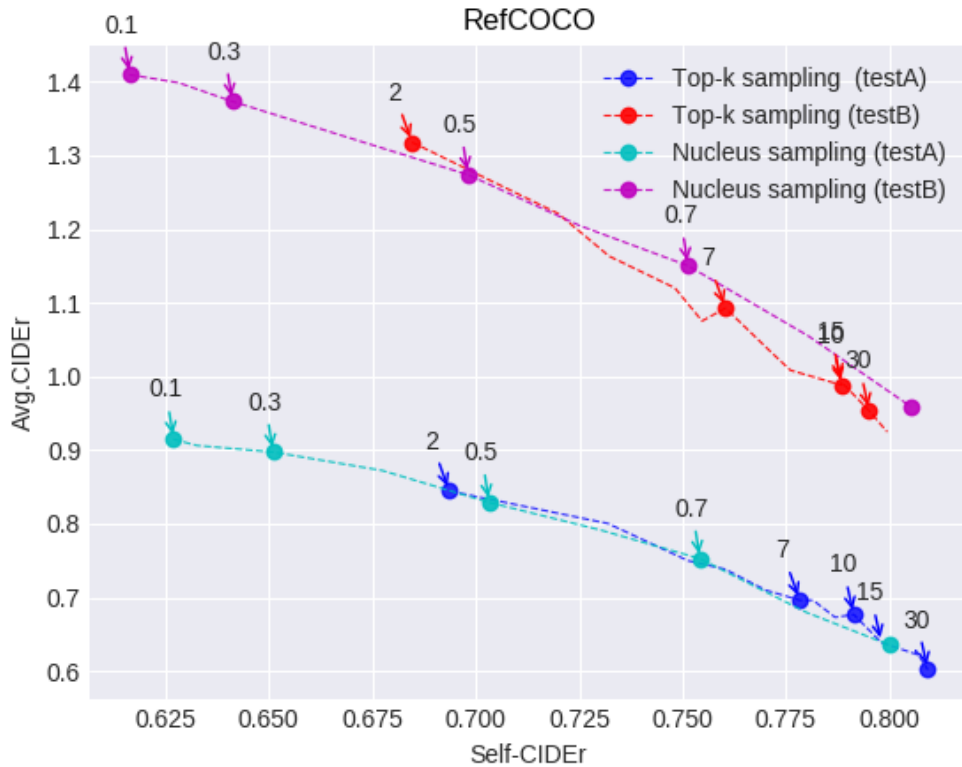


**Figure 5.3:** Self-CIDEr and average CIDEr scores for random sampling with different temperature values for RefCOCO+ dataset. The language model used is the proposed transformer trained with cross-entropy (XE) and fine-tuned with the proposed RL objective.

that have been assigned high-probability, while nucleus sampling keeps those tokens whose cumulative probability mass exceeds a pre-defined threshold  $q$ . Figures 5.4, 5.4 illustrate how the two strategies affect the accuracy-diversity trade-off. It is observed that nucleus sampling achieves higher avg. CIDEr compared to top- $k$  sampling. It is hypothesized that the reason for which top- $k$  results in lower avg. CIDEr scores is that, the distribution is truncated to a fixed number of tokens regardless of the input. There might be cases that there are too many or too few probably tokens. Thus, the fixed number of tokens could potentially lead to sub-optimal solutions. On the contrary, nucleus sampling addresses this issue by dynamically distilling the learned distribution.

## 5.2.4 Maximization-based Decoding Methods

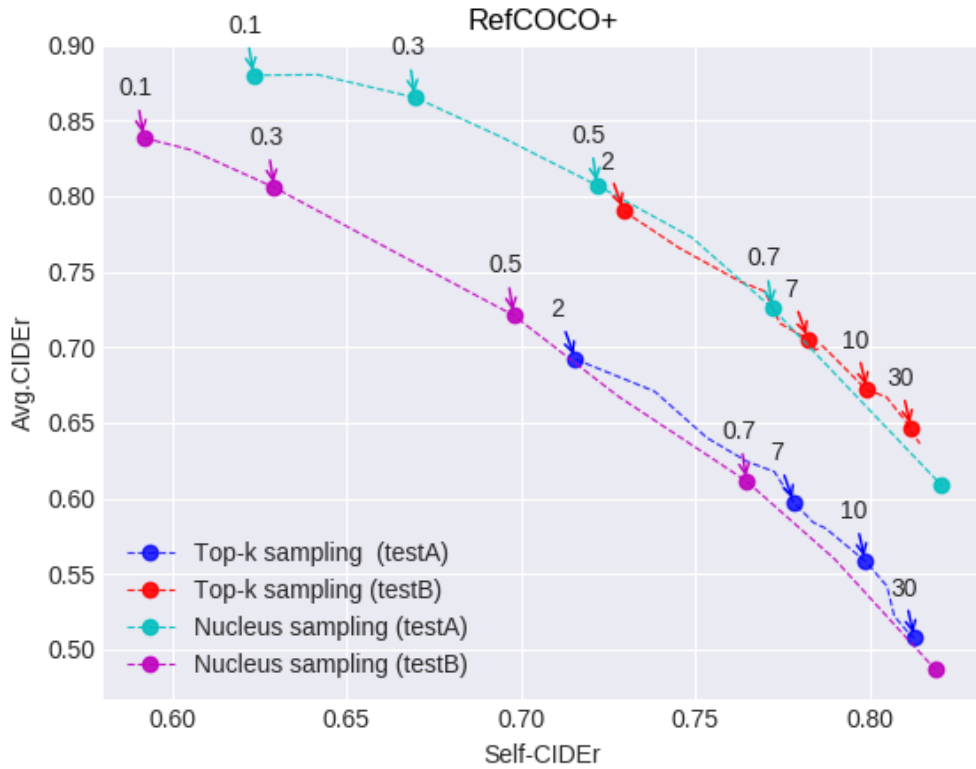
The cross-entropy loss (see Equation 3.18) is minimized when the learned distribution concentrates to the correct ground-truth token. This, ideally, leads to a peaked



**Figure 5.4:** Self-CIDEr and average CIDEr scores for top- $k$  and nucleus sampling for varying  $k$  and  $q$  values for RefCOCO dataset. The temperature was set to  $T = 1$ . The language model used is the proposed transformer trained with cross-entropy.

probability distribution. Hence, the maximization-based decoding methods assume that the model assigns higher probability to higher quality output, and thus they strive to find the sequence with the highest probability tokens. However, the model’s high-confidence over regions of the vocabulary overestimates the use of frequent words resulting in repetition of common words and phrases. Thus, first is investigated how the model’s confidence affects the trade-off between diversity and accuracy for beam search by varying the softmax temperature. Figure 5.7 (top left and right) shows how the temperature modulates the quality-diversity trade-off for RefCOCO dataset. We first observe that unlike the random-based methods, lowering the temperature increases the diversity. As temperature increases ( $\leq 1.5$ ), beam search generates sets with higher average CIDEr. Further increase in temperature ( $> 2$ ) hurts both accuracy and diversity.

Next it is examined how diverse beam search (see Section 5.1.1.3), a diversity-promoting variant of beam search modulates the accuracy-diversity trade-off. The

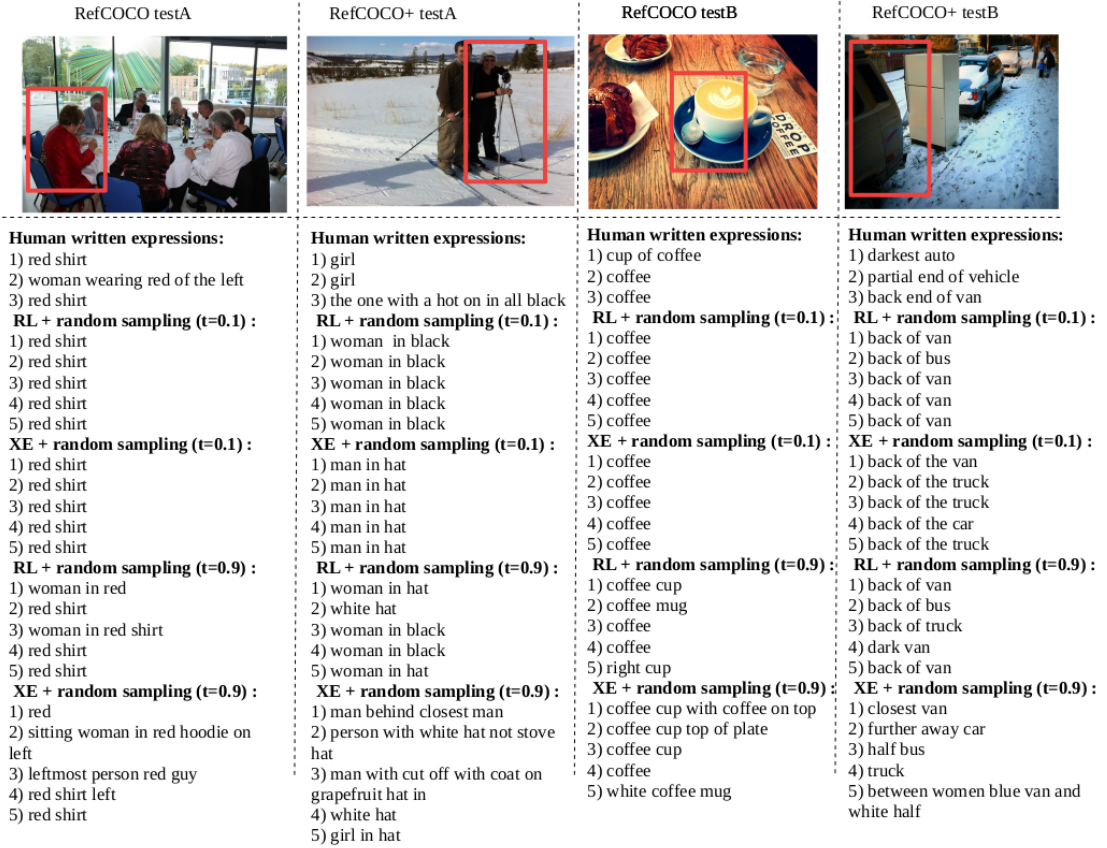


**Figure 5.5:** Self-CIDEr and average CIDEr scores for top- $k$  and nucleus sampling for varying  $k$  and  $q$  values for RefCOCO+ dataset. The temperature was set to  $T = 1$ . The language model used is the proposed transformer trained with cross-entropy.

trade-off is controlled by the diversity strength parameter  $\lambda$ , which we vary between  $[0.1, 2]$ . In Figure 5.7 (bottom left and right) it is observed that as in sampling with temperature, lowering the  $\lambda$  values decreases the diversity. Comparing the DBS with BS with temperature, it is observed that for the same average CIDEr values BS achieves higher diversity.

### 5.2.5 Human evaluation of sets of Referring Expressions

The analysis performed in the previous sections gave vital insights into how the different decoding methods move in the accuracy-diversity space. However, human evaluation is still required to measure the quality and the diversity of the generated expressions. Thus, human evaluation was conducted in order to evaluate the performance of the decoding algorithms along the entire quality-diversity space. In other words, the objective of our human evaluation is to measure the effect that diversity has on the quality

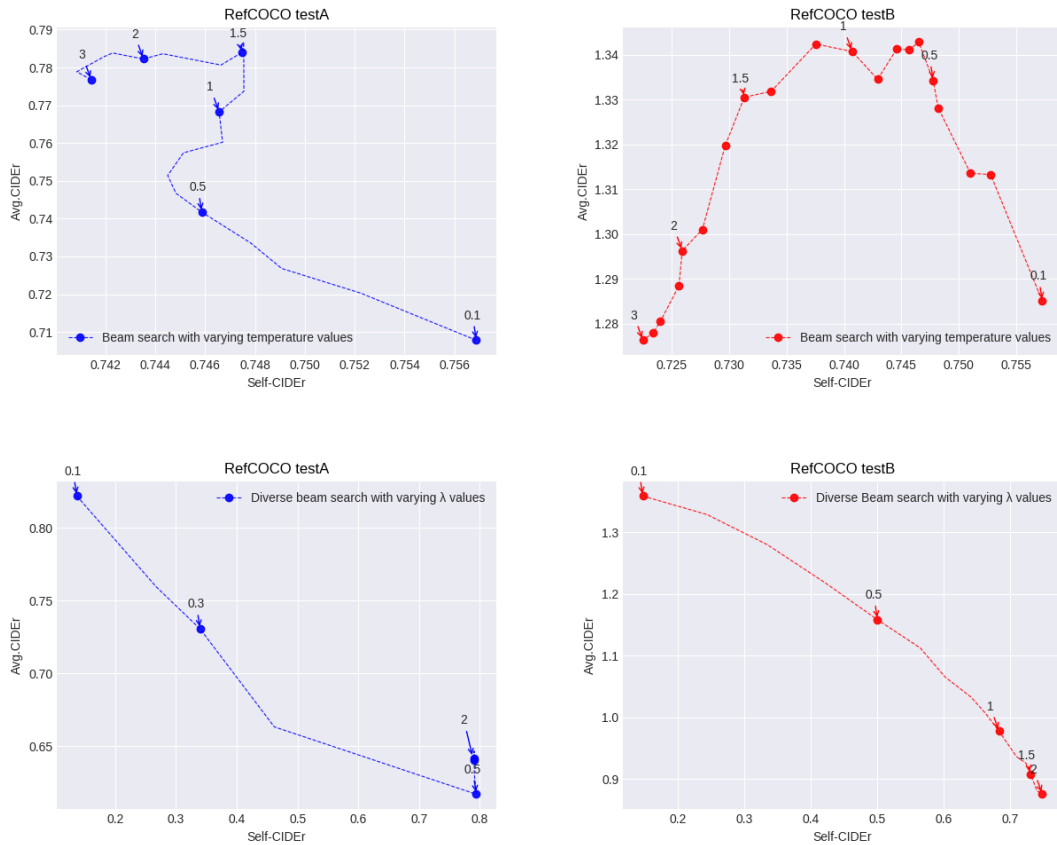


**Figure 5.6:** Examples of objects and sets of expressions drawn from RefCOCO and RefCOCO+ datasets decoded with random sampling with varying temperature values. Human written expressions are also presented.

Random Sampling	$T \in [0.3, 0.5, 0.7, 0.9]$
top- $k$ Sampling	$k \in [2, 5, 10, 15]$
Nucleus Sampling	$q \in [0.3, 0.5, 0.7, 0.9]$
Diverse beam Search	$\lambda \in [0.3, 0.6, 0.9, 1.4]$
Beam search	$T \in [0.1, 1]$

**Table 5.2:** Hyperparameter configurations used in our human evaluation for each of the decoding strategies.

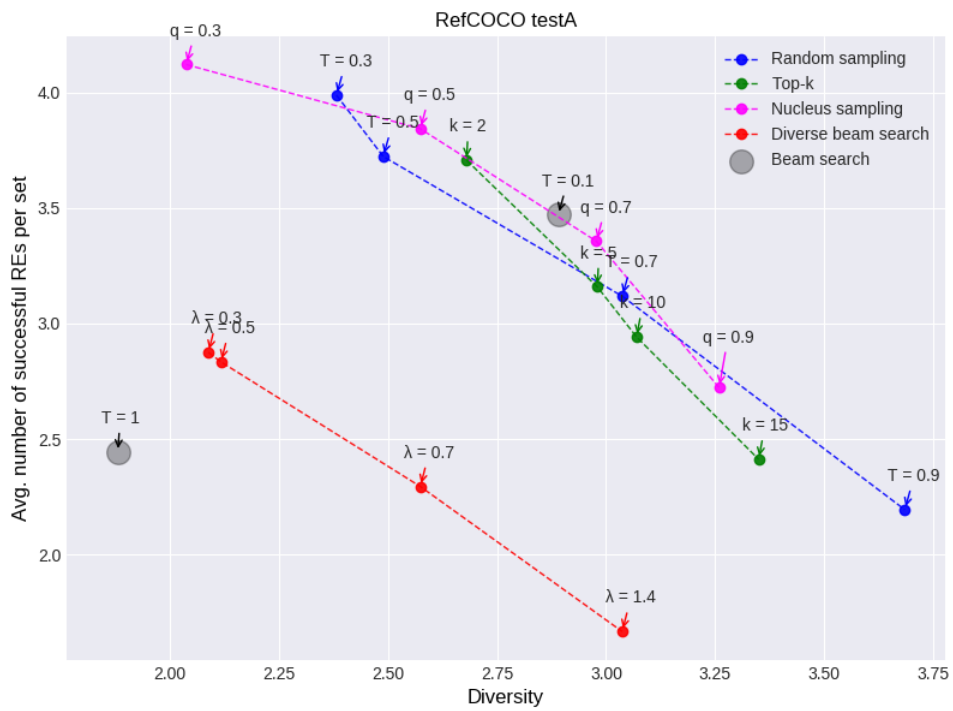
of sets of referring expressions. The decoding strategies used along with the chosen hyperparameters are shown in Table 5.2. The human evaluation protocol is the following. First, 25 objects were selected and for each object a set of 5 expressions was created. Second, each object along with a set of referring expressions was showed to three human annotators. For each of the expressions within a set, human annotators were asked to evaluate whether or not the expression describes the referent object



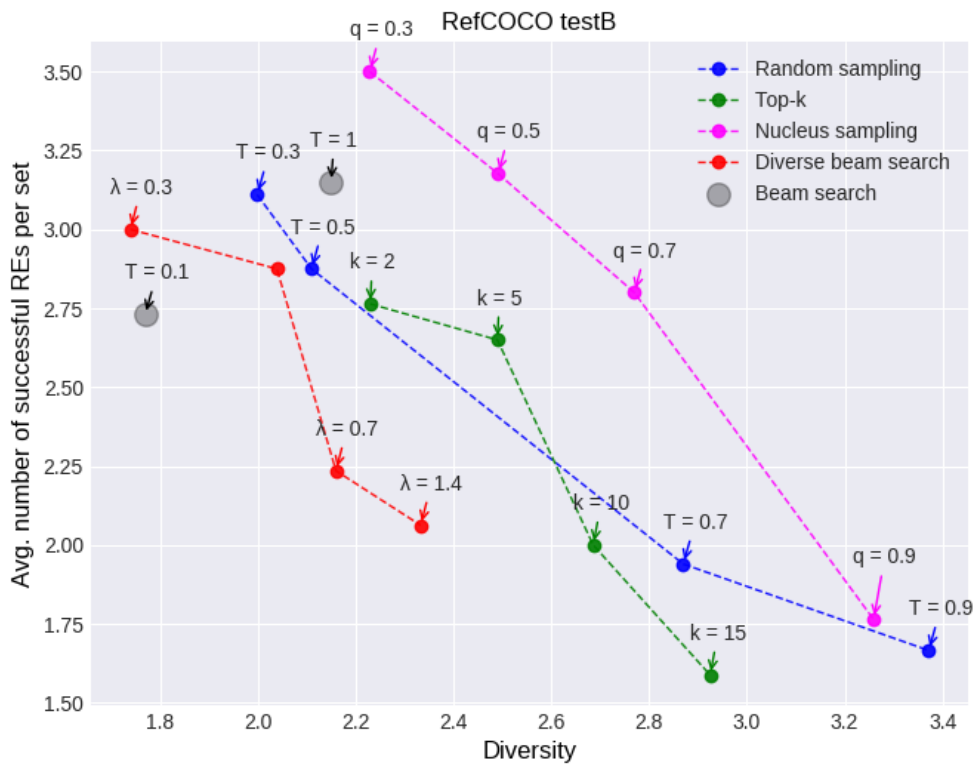
**Figure 5.7:** Self-CIDEr and average CIDEr scores for beam search and diverse beam search with varying temperature and diversity strength values for RefCOCO dataset.

unambiguously. An expression was considered successful if two annotators agreed that the object is described unambiguously by the expression. The quality score of a set is the number of successful referring expressions it contains, while the overall quality score for a hyperparameter configuration is the average number of successful referring expressions of all sets. Furthermore, human annotators were asked to give a diversity score (from 1 to 5, the higher the better) for each set. The diversity score of a set is the average score given by the three human annotators, while the diversity score for each hyperparameter configuration is the average diversity score of all sets.

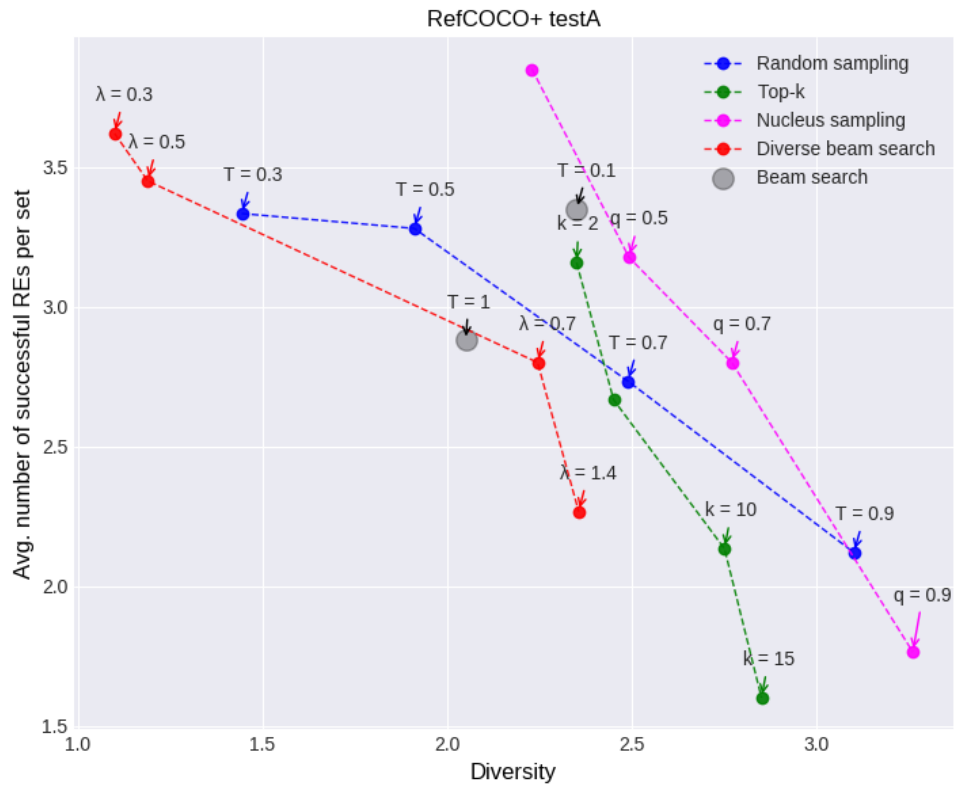
Figures 5.8, 5.9, 5.10 and 5.11 present the results of the human evaluation study. Specifically, beam search and diverse beam search do not produce sets with the highest generation quality. Nucleus sampling with  $q = 0.3$  consistently produces sets that have the highest quality ratings in both datasets. A natural question that arises is why maximization-based algorithms underperform when it comes to the generation of a



**Figure 5.8:** Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO testA.



**Figure 5.9:** Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO testB.

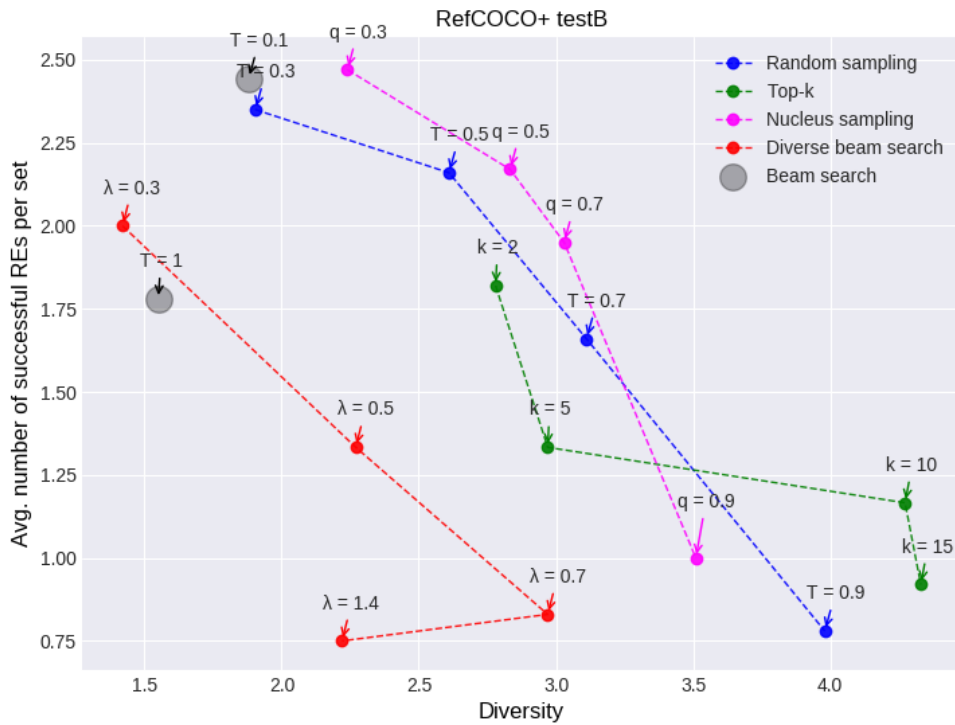


**Figure 5.10:** Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO+ testA.

high quality set. Figure 5.12 shows examples of referent objects and the associated sets of referring expressions for both decoding strategies for different hyperparameters. It is observed that both decoding strategies generate duplicate expressions within a set that contain incorrect or shorter expressions that do not convey enough information to facilitate the identification of the target object. Thus, reducing the overall quality of the set. Furthermore, comparing the default softmax temperature for beam search ( $T = 1$ ) with a sharper distribution ( $T = 0.1$ ), it is observed that the latter produces sets that have higher quality and diversity. One explanation for this behavior is that reducing the temperature, leads to the exploration and expansion of hypotheses that do not stem from one predominant root hypothesis. This is consistent with the examples presented in Figure 5.12.

Furthermore, it is observed that the quality of the sets varies significantly for different levels of diversity for all decoding algorithms. The diversity of the sets when aligned with




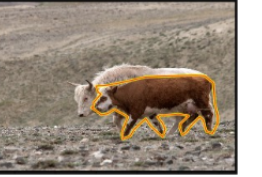




**Figure 5.11:** Human judgment scores for quality and diversity for different hyperparameter configurations for RefCOCO+ testB.

the quality is comparable between all the randomization-based decoding algorithms. It is at the extremes of their hyperparameters range, where the decoding algorithms heavily affect sampling that their performance diverges. Based on the results shown in Figures 5.8,5.9,5.10 and 5.11 the following observations can be made:

- Higher diversity results in lower human judgement scores for quality.
- Nucleus sampling produces sets with higher quality for the same level of diversity between all the decoding strategies, with random sampling performing second best, followed closely by top- $k$  sampling.
- Diverse beam search produces consistently sets with the least diversity.
- Beam search produces higher quality and diversity sets when the softmax temperature is set to  $T = 0.1$  compared to the default value. Interestingly, it produces sets with higher diversity than diverse beam search.

RefCOCO testA	RefCOCO+ testA	RefCOCO testB	RefCOCO+ testB
			
<b>Beam search (<math>t=0.1</math>) :</b> 1) white shirt 2) gray shirt 3) man in gray shirt 4) girl in white 5) girl standing <b>Beam search (<math>t=1</math>) :</b> 1) white shirt 2) white shirt standing 3) man in white 4) gray tshirt 5) gray shirt <b>Diverse beam search (<math>\lambda=0.3</math>) :</b> 1) white shirt 2) white shirt 3) white shirt 4) white shirt 5) standing <b>Diverse beam search (<math>\lambda=0.6</math>) :</b> 1) white shirt 2) white shirt 3) guy standing 4) standing 5) far right person <b>Diverse beam search (<math>\lambda=0.9</math>) :</b> 1) girl in gray 2) gray shirt standing 3) guy standing 4) white shirt standing 5) man on left <b>Diverse beam search (<math>\lambda=1.4</math>) :</b> 1) white shirt 2) gray tshirt 3) white shirt guy 4) man in white 5) far right person	<b>Beam search (<math>t=0.1</math>) :</b> 1) big elephant 2) elephant on left 3) elephant in front 4) front elephant 5) bigger elephant <b>Beam search (<math>t=1</math>) :</b> 1) big elephant 2) front elephant 3) elephant 4) elephant 5) elephant on left <b>Diverse beam search (<math>\lambda=0.3</math>) :</b> 1) big elephant 2) big elephant 3) big elephant 4) elephant 5) elephant <b>Diverse beam search (<math>\lambda=0.6</math>) :</b> 1) big elephant 2) big elephant 3) the big elephant 4) elephant 5) front elephant <b>Diverse beam search (<math>\lambda=0.9</math>) :</b> 1) big elephant 2) big elephant 3) the big elephant 4) elephant 5) front elephant <b>Diverse beam search (<math>\lambda=1.4</math>) :</b> 1) big elephant 2) the big elephant 3) elephant 4) front 5) right	<b>Beam search (<math>t=0.1</math>) :</b> 1) man with glasses 2) black shirt 3) guy with glasses 4) the man with glasses 5) bald guy <b>Beam search (<math>t=1</math>) :</b> 1) man in black 2) man in black 3) black shirt 4) man in black 5) man in black <b>Diverse beam search (<math>\lambda=0.3</math>) :</b> 1) black shirt 2) black shirt 3) black shirt 4) black shirt 5) man with glasses <b>Diverse beam search (<math>\lambda=0.6</math>) :</b> 1) man with glasses 2) black shirt 3) glasses 4) black shirt 5) glasses <b>Diverse beam search (<math>\lambda=0.9</math>) :</b> 1) sunglasses 2) glasses 3) black shirt 4) black shirt 5) glasses <b>Diverse beam search (<math>\lambda=1.4</math>) :</b> 1) sunglasses 2) glasses 3) glasses 4) black shirt 5) guy	<b>Beam search (<math>t=0.1</math>) :</b> 1) brown cow 2) brown cow 3) the one with the white face 4) closest cow 5) closest cow <b>Beam search (<math>t=1</math>) :</b> 1) brown cow 2) brown cow 3) cow 4) cow 5) cow <b>Diverse beam search (<math>\lambda=0.3</math>) :</b> 1) brown cow 2) brown cow 3) brown cow 4) cow 5) cow <b>Diverse beam search (<math>\lambda=0.6</math>) :</b> 1) brown cow 2) brown cow 3) brown cow 4) cow 5) cow <b>Diverse beam search (<math>\lambda=0.9</math>) :</b> 1) cow 2) brown cow 3) brown cow 4) brown 5) brown <b>Diverse beam search (<math>\lambda=1.4</math>) :</b> 1) sunglasses 2) white 3) brown 4) brown 5) light brown cow

**Figure 5.12:** Examples of objects and sets of expressions drawn from RefCOCO and RefCOCO+ datasets. The expressions were decoded with beam search and diverse beam search.

### 5.3 Conclusions

State-of-the-art systems are evaluated to generate a single referring expression. However, multiple referring expressions are potentially correct for a target object- a property that is found in the human written referring expressions. Thus, in this chapter it is attempted to reproduce the diversity found in the human written expressions by generating sets of referring expressions. Specifically, the ability of existing decoding algorithms to generate sets of referring expressions was evaluated by comparing their performance along the entire quality-diversity space. Furthermore, the first large-scale human evaluation study in REG, that compares the quality of sets of referring expressions at the same levels of diversity was introduced. First, it was found that beam search

produces sets with higher quality and diversity when the softmax temperature is set to  $T = 0.1$  compared to the default value  $T = 1$ . Second, both beam search and diverse beam search result in less successful expressions per set compared to the rest decoding algorithms at equal points of diversity. It was showed that duplicate wrong expressions within the sets reduce the quality significantly. Finally, the findings of this chapter suggest that nucleus sampling, produces higher quality sets at the same levels of diversity amongst the compared decoding strategies, with random sampling performing second best, followed by top- $k$  sampling.

## *Conclusions and Future Directions*

---

This thesis has developed and evaluated neural REG approaches that aim to produce: (1) referring expressions that are *unambiguous* and *natural*; and (2) sets of diverse referring expressions. The main contributions and findings are summarized in Section 6.1 and possible avenues for future work are explored in Section 6.2.

### **6.1 Contributions and Findings**

- **RQ1:** *To what extent language models affect the ambiguity and naturalness of referring expressions?*

This thesis initially investigated the effect that language models have on the generation of unambiguous and natural referring expressions. Specifically, two novel language models were developed, evaluated and compared. Thus the following two contributions were made:

- A novel object attention model is proposed that allows the attention mechanism to be calculated at the level of the target object. The object attention mechanism aims at connecting the encoder and decoder, thus aligning better object visual features-to-word interactions. The produced referring expressions were compared with the output of a non-attentive LSTM and the following conclusions can be drawn. First, the expressions produced by the proposed model show an improvement in determining when a relationship

between objects should be expressed, and determining what that relationship should be. Furthermore, the produced referring expressions show an improvement in including appearance and location attributes of the target object. This observation confirms the claims made in this thesis that the object attention mechanism would assist the model in determining both the relationship between objects, but also determine fine appearance details of the target object. Furthermore, significant improvements were noticed in CIDEr and  $BLEU_1$  scores that serve as indication that the produce referring expressions describe the target object with less ambiguity and are similar to those produced by human writers. To validate those claims the output was given to human judges and its was found that the 69.16% of the referring expressions was successfully describing objects of ReFCOCO dataset, while the 95.8% of those were indistinguishable from human written expressions.

- To further demonstrate the benefits of attention in neural REG, a novel transformer-based architecture is proposed. Specifically, a novel layer configuration is proposed in order to provide the network with a global “context” signal by connecting each layer of the encoder with the respective layer of the decoder. The produced referring expressions were compared with the output of those produced by a model following the original architecture. It was found that the proposed model improves over the standard transformer in inferring fine appearance and location attributes of the target object. This observation is in line with the assumption made in this thesis that utilizing features with different degree of modification at each layer, better models the relationships between visual elements and words. Furthermore, for both CIDEr and  $BLEU_1$  the proposed transformer produces higher scores than the standard transformer. To further evaluate whether the produced referring expressions are less ambiguous and indistinguishable from those produced by humans a human evaluation study was conducted. It was found that in terms of task success the proposed model achieves state-of-

the-art results in both datasets. Furthermore, the 97.3% and 94.14% of the expressions generated for objects of ReFCOCO and ReFCOCO+ datasets were indistinguishable from those produced by humans. Finally, when comparing the two proposed models the following conclusions can be drawn: (1) the proposed transformer model is more effective in task success and (2) produces expressions that are less distinguishable from those produced by humans. This is due to the fact that it produces referring expressions that describe the main aspects of the target object more accurately.

- **RQ2** *To what extent training objectives affect the ambiguity, naturalness and diversity of the referring expressions?*

In recent years, encoder-decoder models have gained a lot of popularity and provide state-of-the-art results in a wide variety of tasks such as machine translation, image captioning and text summarization. However such models suffer from two common problems: (1) the exposure bias; and (2) inconsistency between train/test objectives. Recent works address these two problems by leveraging methods from reinforcement learning. A major limitation that stems from those methods is that, the expected gradient exhibits high variance and without proper normalization it leads to unstable training. Thus, this thesis makes the following contribution:

- A novel sequence level optimization approach to REG that is based on the REINFORCE algorithm. The proposed method is compared to the self-critical sequence training, a state-of-the-art optimization technique. It was found that the proposed method results in lower gradient variance for both language models that were tested which indicates the robustness of the approach. This is in line with the assumption made in this thesis that using multiple samples to estimate the expectation will reduce the variance. Furthermore, when a model is optimized with the proposed RL objective achieves higher CIDEr and  $BLEU_1$  scores. To further evaluate the contribution of the proposed optimization approach human evaluation

was conducted. Specifically, the proposed approach improve the results in ReFCOCO testA and testB from 76.95% to 81.66% and from 78.10% to 83.33% respectively, achieving state-of-the-arts results in task success. However, it reduces the naturalness of the generated output and increases the repetitiveness of the generated output.

- To alleviate these issues, this thesis makes a another contribution by proposing a novel strategy for training REG models, using minimum risk training (MRT) with maximum likelihood estimation (MLE). A detailed analysis shows that when a REG model is trained with the proposed approach, uses a larger vocabulary, produces longer referring expressions and generates more uni-grams and bi-grams than a model that is trained with the proposed RL method. Finally, it was shown that combining the RL and MLE objective improves the naturalness and diversity of the produced referring expressions.
- **RQ3** *To what extent decoding methods affect the ambiguity, naturalness and diversity of the referring expressions?*

Choosing the right decoding algorithm is critical in controlling the trade-off between generation quality and diversity. However, there presently exists no consensus on which decoding procedure is best or even the criteria by which to compare them. As presented in Section 2.2, recent efforts have focused primarily on altering the model architecture, visual input and training objectives but there has been significantly less progress towards evaluating improvements in decoder performance. Furthermore, Neural REG approaches are evaluated to produce one referring expression. However, multiple referring expressions are correct for one target object. Thus, this thesis makes a contribution towards this direction by extending the investigation to sets of referring expressions. Furthermore, it presents the first large-scale human evaluation to measure how the hyperparameters of each decoding algorithm, affect the diversity and the quality of sets of referring expressions. Based on the analysis conducted the following conclusion

can be drawn. First, it was found that reducing the softmax temperature to  $T = 0.1$  results in sets with higher quality and diversity. Secondly, beam search and diverse beam search produces sets with lower task success rate due to duplicate wrong expressions within the sets. Finally, it was found that nucleus sampling produces higher quality sets at the same levels of diversity, with random sampling performing second best.

## 6.2 Limitations and Future Work

The work presented in this thesis can be extended as follows:

- **Discriminative attention:** In this thesis the attention mechanism was employed over a set of object features in order to retain richer information that is useful for a more comprehensive referring expression generation. A limitation that stems from this approach is that, it does not explicitly penalise information that is ambiguous. An intuitive extension of this approach is the introduction of a second attention mechanism that attends to objects that are similar to the target object. In other words, the two regions are contrastively attended so that they can be easily discriminated.
- **Reinforcement learning:** Chapter 4 investigates leveraging RL approaches to optimize REG models. It relied on well-defined rewards functions such as BLUE and CIDEr for providing feedback to the model. However, relying on such metrics creates a different set of problems. Specifically, using those functions as reward encourages the model to produce syntactically similar expressions to those produced by humans. Furthermore, an expression that is semantically correct but syntactically varies from the ground truth expressions will be penalized. This problem could be possible addressed by leveraging Inverse Reinforcement Learning (123) approaches that learn their own reward functions.
- **Types of diversity:** This thesis investigated decoding methods for generating diverse sets of referring expressions. The diversity was defined as the difference



between two or more expressions in terms of different words, phrases, sentence structures and semantics. Therefore, further analysis is required to identify the source of diversity.

- **Evaluation** This thesis presented a large scale evaluation study to measure the diversity of the produced referring expressions. However, despite the growing interest in producing diverse output by NLG systems, there is currently no principled method of evaluating the diversity of the output of an NLG system. Therefore, future work should: (1) establish best practices for eliciting diversity human judgments; and (2) introduce a principled and consensual diversity evaluation metric to facilitate the comparison of different approaches.

# References

---

- [1] Peter Anderson et al. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: *CVPR*. 2018 (cit. on pp. 22, 29).
- [2] Peter Anderson et al. “SPICE: Semantic Propositional Image Caption Evaluation”. In: *Computer Vision – ECCV 2016*. Ed. by Bastian Leibe et al. Cham: Springer International Publishing, 2016, pp. 382–398. ISBN: 978-3-319-46454-1 (cit. on p. 20).
- [3] Jimmy Lei Ba, Jamie Ryan Kiros and Geoffrey E. Hinton. *Layer Normalization*. 2016. arXiv: 1607.06450 [stat.ML] (cit. on p. 33).
- [4] Samy Bengio et al. “Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks”. In: *Advances in Neural Information Processing Systems*. 2015 (cit. on pp. 4, 19, 45).
- [5] Bernd Bohnet. *Generation of Referring Expression with an Individual Imprint*. Athens, Greece, 2009. URL: <http://www.aclweb.org/anthology/W09-0631> (cit. on p. 10).
- [6] Bernd Bohnet. “IS-FBN, IS-FBS, IS-IAC: The Adaptation of Two Classic Algorithms for the Generation of Referring Expressions in order to Produce Expressions like Humans Do”. In: 2007 (cit. on p. 10).
- [7] Bernd Bohnet. “The Fingerprint of Human Referring Expressions and Their Surface Realization with Graph Transducers”. In: *Proceedings of the Fifth International Natural Language Generation Conference*. INLG '08. Salt Fork, Ohio:

- 
- Association for Computational Linguistics, 2008, pp. 207–210. URL: <http://dl.acm.org/citation.cfm?id=1708322.1708366> (cit. on p. 11).
- [8] Thiago Castro Ferreira et al. “Neural data-to-text generation: A comparison between pipeline and end-to-end architectures”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 552–562. DOI: 10.18653/v1/D19-1052. URL: <https://www.aclweb.org/anthology/D19-1052> (cit. on pp. 1, 26).
- [9] Long Chen et al. “SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning.” In: *CoRR* abs/1611.05594 (2016). URL: <http://dblp.uni-trier.de/db/journals/corr/corr1611.html#ChenZXNSC16> (cit. on p. 22).
- [10] Kyunghyun Cho. “Noisy Parallel Approximate Decoding for Conditional Recurrent Language Model”. In: *CoRR* abs/1605.03835 (2016). arXiv: 1605.03835. URL: <http://arxiv.org/abs/1605.03835> (cit. on pp. 16, 65).
- [11] Kyunghyun Cho et al. “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”. In: *CoRR* abs/1406.1078 (2014). arXiv: 1406.1078. URL: <http://arxiv.org/abs/1406.1078> (cit. on p. 26).
- [12] Kyunghyun Cho et al. “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches”. In: *Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*. 2014 (cit. on p. 1).
- [13] T. Cover and P. Hart. “Nearest Neighbor Pattern Classification”. In: *IEEE Trans. Inf. Theor.* 13.1 (Sept. 2006), pp. 21–27. ISSN: 0018-9448. DOI: 10.1109/TIT.1967.1053964. URL: <http://dx.doi.org/10.1109/TIT.1967.1053964> (cit. on p. 10).
- [14] Robert Dale. “Cooking up Referring Expressions”. In: *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics*. 1989 (cit. on p. 9).
-

- 
- [15] Robert Dale and Ehud Reiter. “Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions”. In: *Cognitive Science* 19.2 (1995), pp. 233–263. ISSN: 1551-6709. DOI: 10.1207/s15516709cog1902\_3. URL: [http://dx.doi.org/10.1207/s15516709cog1902\\_3](http://dx.doi.org/10.1207/s15516709cog1902_3) (cit. on p. 9).
- [16] Robert Dale and Ehud Reiter. “Computational interpretations of the Gricean maxims in the generation of referring expressions”. English. In: *Cognitive Science* 19.2 (Apr. 1995), pp. 233–263. ISSN: 0364-0213. DOI: 10.1016/0364-0213(95)90018-7 (cit. on p. 9).
- [17] Kees van Deemter et al. “Generation of Referring Expressions: Assessing the Incremental Algorithm”. In: *Cognitive Science* 36.5 (2012), pp. 799–836. ISSN: 1551-6709. DOI: 10.1111/j.1551-6709.2011.01205.x. URL: <http://dx.doi.org/10.1111/j.1551-6709.2011.01205.x> (cit. on p. 12).
- [18] Giuseppe Di Fabbrizio, Amanda J. Stent and Srinivas Bangalore. “Referring Expression Generation Using Speaker-based Attribute Selection and Trainable Realization (ATTR)”. In: *Proceedings of the Fifth International Natural Language Generation Conference*. INLG ’08. Salt Fork, Ohio: Association for Computational Linguistics, 2008, pp. 211–214. URL: <http://dl.acm.org/citation.cfm?id=1708322.1708367> (cit. on p. 11).
- [19] Angela Fan, Mike Lewis and Yann Dauphin. “Hierarchical Neural Story Generation”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018 (cit. on pp. 1, 18, 27, 67).
- [20] Thiago Ferreira and Ivandre Paraboni. “Generating natural language descriptions using speaker-dependent information”. In: *Natural Language Engineering* 23 (Feb. 2017), pp. 1–22. DOI: 10.1017/S1351324917000079 (cit. on p. 12).
- [21] Jessica Fidler and Yoav Goldberg. “Controlling Linguistic Style Aspects in Neural Language Generation”. In: *Proceedings of the Workshop on Stylistic Variation*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 94–104. DOI: 10.18653/v1/W17-4912. URL: <https://www.aclweb.org/anthology/W17-4912> (cit. on p. 66).
-

- 
- [22] Albert Gatt, Ielka van der Sluis and Kees van Deemter. “Evaluating Algorithms for the Generation of Referring Expressions Using a Balanced Corpus”. In: *Proceedings of the Eleventh European Workshop on Natural Language Generation*. ENLG '07. Germany: Association for Computational Linguistics, 2007, pp. 49–56. URL: <http://dl.acm.org/citation.cfm?id=1610163.1610172> (cit. on p. 12).
- [23] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. The MIT Press, 2016. ISBN: 0262035618 (cit. on p. 66).
- [24] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani et al. Vol. 27. Curran Associates, Inc., 2014. URL: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf> (cit. on p. 15).
- [25] Han Guo, Ramakanth Pasunuru and Mohit Bansal. “Soft Layer-Specific Multi-Task Summarization with Entailment and Question Generation”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2018 (cit. on pp. 1, 27).
- [26] Longteng Guo et al. “Normalized and Geometry-Aware Self-Attention Network for Image Captioning”. In: *CoRR* abs/2003.08897 (2020). arXiv: 2003.08897. URL: <https://arxiv.org/abs/2003.08897> (cit. on p. 24).
- [27] Di He et al. “Decoding with Value Networks for Neural Machine Translation”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: <https://proceedings.neurips.cc/paper/2017/file/2b24d495052a8ce66358eb576b8912c8-Paper.pdf> (cit. on p. 20).
- [28] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: <http://arxiv.org/abs/1512.03385> (cit. on p. 26).

- 
- [29] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on pp. 19, 35).
- [30] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90. URL: <https://doi.org/10.1109/CVPR.2016.90> (cit. on p. 33).
- [31] Simao Herdade et al. “Image Captioning: Transforming Objects into Words”. In: *CoRR* abs/1906.05963 (2019). arXiv: 1906.05963. URL: <http://arxiv.org/abs/1906.05963> (cit. on p. 24).
- [32] Raquel Hervás, Virginia Francisco and Pablo Gervás. “Assessing the Influence of Personal Preferences on the Choice of Vocabulary for Natural Language Generation”. In: *Inf. Process. Manage.* 49.4 (July 2013), pp. 817–832. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2013.01.006. URL: <http://dx.doi.org/10.1016/j.ipm.2013.01.006> (cit. on p. 12).
- [33] Raquel Hervás et al. “Influence of personal choices on lexical variability in referring expressions”. In: *Natural Language Engineering* 22.02 (2016), pp. 257–290. ISSN: 1469-8110. DOI: <https://doi.org/10.1017/S1351324915000182> (cit. on p. 13).
- [34] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Comput.* 9.8 (Nov. 1997), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. URL: <https://doi.org/10.1162/neco.1997.9.8.1735> (cit. on p. 1).
- [35] Sepp Hochreiter and Jürgen Schmidhuber. “Long Short-Term Memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780 (cit. on p. 13).
- [36] Ari Holtzman et al. “Learning to Write with Cooperative Discriminators”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018 (cit. on pp. 1, 18, 27).
-

- 
- [37] Ari Holtzman et al. “The Curious Case of Neural Text Degeneration”. In: *International Conference on Learning Representations*. 2020. URL: <https://openreview.net/forum?id=rygGQyrFvH> (cit. on pp. 16, 18, 63, 66, 67).
- [38] Lun Huang et al. “Adaptively Aligned Image Captioning via Adaptive Attention Time”. In: *CoRR* abs/1909.09060 (2019). arXiv: 1909.09060. URL: <http://arxiv.org/abs/1909.09060> (cit. on p. 23).
- [39] Lun Huang et al. “Attention on Attention for Image Captioning”. In: *CoRR* abs/1908.06954 (2019). arXiv: 1908.06954. URL: <http://arxiv.org/abs/1908.06954> (cit. on p. 24).
- [40] Daphne Ippolito et al. “Comparison of Diverse Decoding Methods from Conditional Language Models”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019 (cit. on p. 16).
- [41] Daphne Ippolito et al. “Comparison of Diverse Decoding Methods from Conditional Language Models”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019 (cit. on p. 68).
- [42] L. C. Jain and L. R. Medsker. *Recurrent Neural Networks: Design and Applications*. 1st. Boca Raton, FL, USA: CRC Press, Inc., 1999. ISBN: 0849371813 (cit. on p. 1).
- [43] Sahar Kazemzadeh et al. “ReferItGame: Referring to Objects in Photographs of Natural Scenes”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014 (cit. on p. 37).
- [44] Mert Kilickaya et al. “Re-evaluating Automatic Metrics for Image Captioning”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 199–209. URL: <https://www.aclweb.org/anthology/E17-1019> (cit. on p. 36).
- [45] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd Inter-
-

- 
- national Conference for Learning Representations, San Diego, 2015. 2014. URL: <http://arxiv.org/abs/1412.6980> (cit. on p. 54).
- [46] Philipp Koehn. “Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models”. In: *Machine Translation: From Real Users to Research*. Ed. by Robert E. Frederking and Kathryn B. Taylor. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 115–124. ISBN: 978-3-540-30194-3 (cit. on p. 63).
- [47] Wouter Kool, Herke van Hoof and Max Welling. “Buy 4 REINFORCE Samples, Get a Baseline for Free!” In: *Deep Reinforcement Learning Meets Structured Prediction, ICLR*. 2019 (cit. on p. 50).
- [48] Emiel Krahmer and Kees van Deemter. “Computational Generation of Referring Expressions: A Survey”. In: *Comput. Linguist.* 38.1 (2012), pp. 173–218 (cit. on pp. 1, 8).
- [49] Alex Krizhevsky, Ilya Sutskever and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Proceedings of the 25th Conference on Advances in Neural Information Processing Systems*. 2012 (cit. on p. 1).
- [50] Guang Li et al. “Entangled Transformer for Image Captioning”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2019 (cit. on p. 24).
- [51] Jiwei Li and Dan Jurafsky. “Mutual Information and Diverse Decoding Improve Neural Machine Translation.” In: *CoRR* abs/1601.00372 (2016). URL: <http://dblp.uni-trier.de/db/journals/corr/corr1601.html#LiJ16> (cit. on pp. 6, 63).
- [52] Jiwei Li and Dan Jurafsky. “Mutual Information and Diverse Decoding Improve Neural Machine Translation.” In: *CoRR* abs/1601.00372 (2016). URL: <http://dblp.uni-trier.de/db/journals/corr/corr1601.html#LiJ16> (cit. on p. 17).
-



- 
- [53] Jiwei Li, Will Monroe and Dan Jurafsky. “Learning to Decode for Future Success”. In: *CoRR* abs/1701.06549 (2017). arXiv: 1701.06549. URL: <http://arxiv.org/abs/1701.06549> (cit. on p. 17).
- [54] Jiwei Li, Will Monroe and Dan Jurafsky. “Learning to Decode for Future Success”. In: *ArXiv* abs/1701.06549 (2017) (cit. on p. 20).
- [55] Jiwei Li et al. “Deep Reinforcement Learning for Dialogue Generation”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2016 (cit. on pp. 1, 27).
- [56] Piji Li, Lidong Bing and Wai Lam. “Actor-Critic based Training Framework for Abstractive Summarization”. In: *CoRR* abs/1803.11070 (2018). arXiv: 1803.11070. URL: <http://arxiv.org/abs/1803.11070> (cit. on p. 20).
- [57] Timothy P. Lillicrap et al. “Continuous control with deep reinforcement learning.” In: *ICLR*. Ed. by Yoshua Bengio and Yann LeCun. 2016. URL: <http://dblp.uni-trier.de/db/conf/iclr/iclr2016.html#LillicrapHPHETS15> (cit. on p. 51).
- [58] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *Computer Vision – ECCV*. Ed. by David Fleet et al. 2014 (cit. on p. 37).
- [59] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context”. In: *CoRR* abs/1405.0312 (2014). arXiv: 1405.0312. URL: <http://arxiv.org/abs/1405.0312> (cit. on p. 17).
- [60] J. Liu, L. Wang and M. H. Yang. “Referring Expression Generation and Comprehension via Attributes”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 4866–4874. DOI: 10.1109/ICCV.2017.520 (cit. on p. 12).
- [61] Jingyu Liu, Liang Wang and Ming-Hsuan Yang. “Referring Expression Generation and Comprehension via Attributes”. In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017 (cit. on p. 25).
-

- 
- [62] Siqi Liu et al. “Improved Image Captioning via Policy Gradient optimization of SPIDeR”. In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 873–881. DOI: 10.1109/ICCV.2017.100. URL: <https://doi.org/10.1109/ICCV.2017.100> (cit. on p. 20).
- [63] Ruotian Luo and Gregory Shakhnarovich. “Comprehension-guided referring expressions”. In: *arXiv preprint arXiv:1701.03439* (2017) (cit. on pp. 2, 15, 21, 25).
- [64] Ruotian Luo and Gregory Shakhnarovich. “Comprehension-guided referring expressions”. In: *CoRR abs/1701.03439* (2017). arXiv: 1701.03439. URL: <http://arxiv.org/abs/1701.03439> (cit. on p. 13).
- [65] Chris J. Maddison, Andriy Mnih and Yee Whye Teh. “The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables”. In: *CoRR abs/1611.00712* (2016) (cit. on p. 56).
- [66] J. Mao et al. “Generation and Comprehension of Unambiguous Object Descriptions”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016 (cit. on p. 35).
- [67] Junhua Mao et al. “Generation and Comprehension of Unambiguous Object Descriptions”. In: *CoRR abs/1511.02283* (2015). arXiv: 1511.02283. URL: <http://arxiv.org/abs/1511.02283> (cit. on pp. 1, 12–15, 21, 36, 54).
- [68] Emiel van Miltenburg, Desmond Elliott and Piek Vossen. “Measuring the Diversity of Automatic Image Descriptions”. In: *Proceedings of the 27th International Conference on Computational Linguistics (ACL)*. 2018 (cit. on p. 54).
- [69] Margaret Mitchell et al. “Midge: Generating Image Descriptions From Computer Vision Detections”. In: *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 747–756. URL: <https://www.aclweb.org/anthology/E12-1076> (cit. on p. 37).
-

- 
- [70] Andriy Mnih and Danilo J. Rezende. “Variational Inference for Monte Carlo Objectives”. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning*. 2016 (cit. on p. 50).
- [71] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *Nature* 518.7540 (Feb. 2015), pp. 529–533. ISSN: 00280836. URL: <http://dx.doi.org/10.1038/nature14236> (cit. on p. 51).
- [72] Shashi Narayan and Claire Gardent. “Deep Learning Approaches to Text Production”. In: *Synthesis Lectures on Human Language Technologies* 13.1 (2020), pp. 1–199. DOI: 10.2200/S00979ED1V01Y201912HLT044. eprint: <https://doi.org/10.2200/S00979ED1V01Y201912HLT044>. URL: <https://doi.org/10.2200/S00979ED1V01Y201912HLT044> (cit. on p. 1).
- [73] Nikolaos Panagiaris, Emma Hart and Dimitra Gkatzia. “Generating unambiguous and diverse referring expressions”. In: *Computer Speech Language* 68 (2021), p. 101184. ISSN: 0885-2308. DOI: <https://doi.org/10.1016/j.csl.2020.101184>. URL: <https://www.sciencedirect.com/science/article/pii/S0885230820301170> (cit. on pp. 28, 46, 64).
- [74] Nikolaos Panagiaris, Emma Hart and Dimitra Gkatzia. “Improving the Naturalness and Diversity of Referring Expression Generation models using Minimum Risk Training”. In: *Proceedings of the 13th International Conference on Natural Language Generation*. Dublin, Ireland: Association for Computational Linguistics, Dec. 2020, pp. 41–51. URL: <https://aclanthology.org/2020.inlg-1.7> (cit. on p. 47).
- [75] Th. Pechmann. “Incremental speech production and referential overspecification”. In: *Linguistics* 27 (1989), pp. 89–110 (cit. on p. 9).
- [76] Marco Pedersoli et al. “Areas of Attention for Image Captioning”. In: *CoRR* abs/1612.01033 (2016). arXiv: 1612.01033. URL: <http://arxiv.org/abs/1612.01033> (cit. on p. 23).

- 
- [77] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN: 1-55860-238-0 (cit. on p. 11).
- [78] Vasili Ramanishka et al. “Top-down Visual Saliency Guided by Captions”. In: *CoRR* abs/1612.07360 (2016). arXiv: 1612.07360. URL: <http://arxiv.org/abs/1612.07360> (cit. on p. 22).
- [79] Marc’Aurelio Ranzato et al. “Sequence Level Training with Recurrent Neural Networks”. In: *Proceedings of the 4th International Conference on Learning Representations ICLR*. 2016 (cit. on pp. 4, 19, 20, 45–47).
- [80] Zhou Ren et al. “Deep Reinforcement Learning-Based Image Captioning with Embedding Reward”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1151–1159. DOI: 10.1109/CVPR.2017.128 (cit. on p. 20).
- [81] Steven J. Rennie et al. “Self-Critical Sequence Training for Image Captioning”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017 (cit. on pp. 4, 20, 21, 25, 45, 49).
- [82] Stephane Ross, Geoffrey Gordon and Drew Bagnell. “A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning”. In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Geoffrey Gordon, David Dunson and Miroslav Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, Nov. 2011, pp. 627–635. URL: <https://proceedings.mlr.press/v15/ross11a.html> (cit. on p. 19).
- [83] John Schulman et al. “High-Dimensional Continuous Control Using Generalized Advantage Estimation”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*. 2016 (cit. on pp. 20, 45).
- [84] Shiqi Shen et al. “Minimum Risk Training for Neural Machine Translation”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computa-
-

- 
- tional Linguistics, Aug. 2016, pp. 1683–1692. DOI: 10.18653/v1/P16-1159. URL: <https://www.aclweb.org/anthology/P16-1159> (cit. on pp. 5, 46).
- [85] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *CoRR* abs/1409.1556 (2014). arXiv: 1409.1556. URL: <http://arxiv.org/abs/1409.1556> (cit. on p. 26).
- [86] Yusuke Sugano and Andreas Bulling. “Seeing with Humans: Gaze-Assisted Neural Image Captioning”. In: *CoRR* abs/1608.05203 (2016). arXiv: 1608.05203. URL: <http://arxiv.org/abs/1608.05203> (cit. on p. 22).
- [87] Ilya Sutskever, Oriol Vinyals and Quoc V Le. “Sequence to Sequence Learning with Neural Networks”. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems*. 2014 (cit. on p. 1).
- [88] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. Second. The MIT Press, 2018. URL: <http://incompleteideas.net/book/the-book-2nd.html> (cit. on pp. 4, 45).
- [89] Christian Szegedy et al. “Going Deeper with Convolutions”. In: *CoRR* abs/1409.4842 (2014). arXiv: 1409.4842. URL: <http://arxiv.org/abs/1409.4842> (cit. on p. 26).
- [90] Kai Sheng Tai, Richard Socher and Christopher D. Manning. “Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks”. In: *CoRR* abs/1503.00075 (2015). arXiv: 1503.00075. URL: <http://arxiv.org/abs/1503.00075> (cit. on p. 24).
- [91] Jiwei Tan, Xiaojun Wan and Jianguo Xiao. “Abstractive Document Summarization with a Graph-Based Attentional Neural Model”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2017 (cit. on pp. 1, 27).
- [92] Hamed Rezazadegan Tavakoli et al. “Can Saliency Information Benefit Image Captioning Models?” In: *CoRR* abs/1704.07434 (2017). arXiv: 1704.07434. URL: <http://arxiv.org/abs/1704.07434> (cit. on p. 22).
-

- 
- [93] Jörg Tiedemann. “News from OPUS — A collection of multilingual parallel corpora with tools and interfaces”. In: 2009 (cit. on p. 18).
- [94] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., 2017, pp. 5998–6008. URL: <http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf> (cit. on pp. 4, 24, 27, 32, 40, 41).
- [95] Ramakrishna Vedantam, C. Lawrence Zitnick and Devi Parikh. “CIDEr: Consensus-based image description evaluation.” In: *CVPR*. 2015 (cit. on p. 20).
- [96] Arun Venkatraman, Martial Hebert and J. Andrew Bagnell. “Improving Multi-Step Prediction of Learned Time Series Models”. In: *AAAI’15*. Austin, Texas: AAAI Press, 2015. ISBN: 0262511290 (cit. on p. 19).
- [97] Jette Viethen and Robert Dale. “Speaker-Dependent Variation in Content Selection for Referring Expression Generation”. In: *Proceedings of the Australasian Language Technology Association Workshop 2010*. 2010 (cit. on p. 2).
- [98] Jette Viethen and Robert Dale. “Speaker-dependent variation in content selection for referring expression generation”. In: *In Proceedings of the Australasian Language Technology Workshop*. 2010 (cit. on pp. 11, 12).
- [99] Jette Viethen, Margaret Mitchell and Emiel Kraahmer. *Graphs and Spatial Relations in the Generation of Referring Expressions*. 2013 (cit. on p. 11).
- [100] Ashwin K. Vijayakumar et al. “Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models”. In: *CoRR* abs/1610.02424 (2016). arXiv: 1610.02424. URL: <http://arxiv.org/abs/1610.02424> (cit. on pp. 16, 17, 63, 65, 66).
- [101] Ashwin K. Vijayakumar et al. “Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models”. In: *CoRR* abs/1610.02424 (2016). arXiv: 1610.02424. URL: <http://arxiv.org/abs/1610.02424> (cit. on pp. 63, 66).
-

- 
- [102] O. Vinyals et al. “Show and tell: A neural image caption generator”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015 (cit. on pp. 1, 26, 27).
- [103] Oriol Vinyals and Quoc V. Le. “A Neural Conversational Model”. In: *CoRR* abs/1506.05869 (2015). URL: <http://arxiv.org/abs/1506.05869> (cit. on pp. 1, 27).
- [104] Qingzhong Wang and Antoni B. Chan. “Describing Like Humans: On Diversity in Image Captioning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on pp. 2, 63, 69).
- [105] Qingzhong Wang and Antoni B. Chan. “Describing Like Humans: On Diversity in Image Captioning”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019 (cit. on p. 5).
- [106] Ronald J. Williams. “Simple statistical gradient-following algorithms for connectionist reinforcement learning”. In: *Machine Learning*. 1992, pp. 229–256 (cit. on p. 49).
- [107] Yonghui Wu et al. “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”. In: *CoRR* abs/1609.08144 (2016). URL: <http://arxiv.org/abs/1609.08144> (cit. on p. 46).
- [108] Kelvin Xu et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. In: *Proceedings of the 32nd International Conference on Machine Learning*. 2015 (cit. on pp. 1, 27).
- [109] Kelvin Xu et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 2048–2057. URL: <https://proceedings.mlr.press/v37/xuc15.html> (cit. on pp. 21, 22).
- [110] Xu Yang, Hanwang Zhang and Jianfei Cai. “Learning to Collocate Neural Modules for Image Captioning”. In: *CoRR* abs/1904.08608 (2019). arXiv: 1904.08608. URL: <http://arxiv.org/abs/1904.08608> (cit. on p. 24).
-

- 
- [111] Xu Yang et al. “Auto-Encoding Scene Graphs for Image Captioning”. In: *CoRR* abs/1812.02378 (2018). arXiv: 1812.02378. URL: <http://arxiv.org/abs/1812.02378> (cit. on p. 23).
- [112] Ting Yao et al. “Exploring Visual Relationship for Image Captioning”. In: *CoRR* abs/1809.07041 (2018). arXiv: 1809.07041. URL: <http://arxiv.org/abs/1809.07041> (cit. on p. 23).
- [113] Ting Yao et al. “Hierarchy Parsing for Image Captioning”. In: *CoRR* abs/1909.03918 (2019). arXiv: 1909.03918. URL: <http://arxiv.org/abs/1909.03918> (cit. on p. 23).
- [114] Licheng Yu et al. “A Joint Speaker-Listener-Reinforcer Model for Referring Expressions”. In: *CoRR* abs/1612.09542 (2016). arXiv: 1612.09542. URL: <http://arxiv.org/abs/1612.09542> (cit. on p. 15).
- [115] Licheng Yu et al. “A Joint Speaker-Listener-Reinforcer Model for Referring Expressions”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition CVPR*. 2017 (cit. on pp. 1, 2, 13, 25, 26, 36, 43, 44, 46).
- [116] Licheng Yu et al. “Modeling Context in Referring Expressions”. In: *Proceedings of the 14th European Conference on Computer Vision (ECCV)*. 2016 (cit. on pp. 1, 2, 13, 18, 21, 25, 26, 36, 37, 54).
- [117] Licheng Yu et al. “Modeling Context in Referring Expressions”. In: *CoRR* abs/1608.00272 (2016). arXiv: 1608.00272. URL: <http://arxiv.org/abs/1608.00272> (cit. on p. 14).
- [118] Wojciech Zaremba and Ilya Sutskever. “Reinforcement Learning Neural Turing Machines”. In: *CoRR* abs/1505.00521 (2015). URL: <http://arxiv.org/abs/1505.00521> (cit. on pp. 20, 45, 48).
- [119] Sina Zarrieß and David Schlangen. “Decoding Strategies for Neural Referring Expression Generation”. In: *Proceedings of the 11th International Conference on Natural Language Generation*. Tilburg University, The Netherlands: Association for Computational Linguistics, Nov. 2018, pp. 503–512. DOI: 10.18653/v1/W18-
-



- 
6563. URL: <https://www.aclweb.org/anthology/W18-6563> (cit. on pp. 1, 13, 16, 18, 26, 36, 54).
- [120] Sina Zarrieß and David Schlangen. “Easy Things First: Installments Improve Referring Expression Generation for Objects in Photographs”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 610–620. DOI: 10.18653/v1/P16-1058. URL: <https://www.aclweb.org/anthology/P16-1058> (cit. on p. 36).
- [121] Sina Zarrieß, Henrik Voigt and Simeon Schüz. “Decoding Methods in Neural Language Generation: A Survey”. In: *Information* 12.9 (2021). ISSN: 2078-2489. DOI: 10.3390/info12090355. URL: <https://www.mdpi.com/2078-2489/12/9/355> (cit. on pp. 64, 65).
- [122] Zheng-Jun Zha et al. “Context-Aware Visual Policy Network for Fine-Grained Image Captioning”. In: *CoRR* abs/1906.02365 (2019). arXiv: 1906.02365. URL: <http://arxiv.org/abs/1906.02365> (cit. on p. 23).
- [123] Brian D. Ziebart et al. “Maximum Entropy Inverse Reinforcement Learning”. In: *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3. AAAI’08*. Chicago, Illinois: AAAI Press, 2008, pp. 1433–1438. ISBN: 9781577353683 (cit. on p. 85).

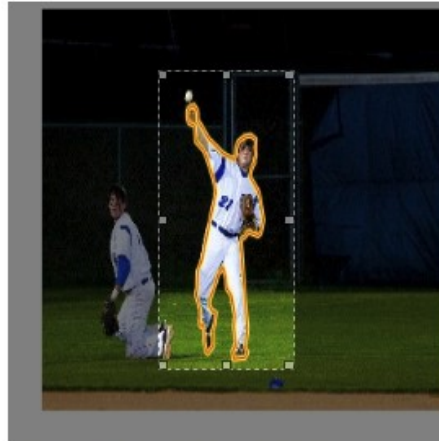
## *Human evaluation questionnaires*

---

All Human judgements were collected using Amazon Mechanical Turk crowdsourcing platform by English native-speaking workers that were specifically qualified for this task. Figure [A.1](#) shows an example question for evaluating the quality of a single referring expression. Figure [A.2](#) shows an example question for evaluating the quality and diversity of a set of referring expressions.

# Referring expression comprehension.

## 1. Question



Draw a box around the object that is described by the following expression:  
*The man with the blue jacket*

It sounds like a person wrote this description.  
*The man with the blue jacket*

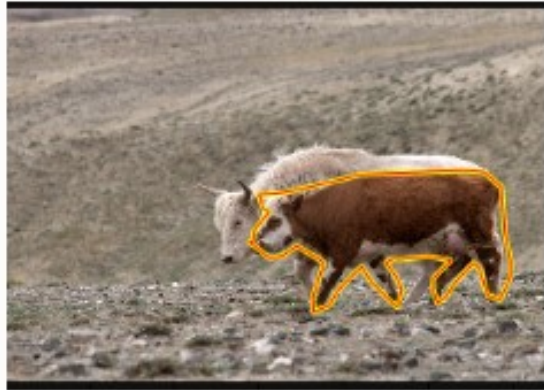
- Yes  
 No

How much do you agree with the following statements?

	Strongly agree		Neither agree nor disagree		Strongly disagree
	1	2	3	4	5
This description is grammatically correct.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This description describes the main attributes of the object incorrectly.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This description does not include extraneous or incorrect information.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
This description refers to more than 2 objects.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The attributes of the highlighted object are mentioned in a reasonable order.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**Figure A.1:** Amazon mechanical turk example question for evaluating the quality of a single referring expression.

## Evaluating object descriptions



If you were asked to choose the object that is described by the sentence, would you have chosen the highlighted object?

	Yes	No
brown cow	<input type="radio"/>	<input type="radio"/>
brown cow	<input type="radio"/>	<input type="radio"/>
cow	<input type="radio"/>	<input type="radio"/>
cow	<input type="radio"/>	<input type="radio"/>
brown and white cow	<input type="radio"/>	<input type="radio"/>

How diverse are the descriptions presented in previous question?

- 5- Very Diverse (The descriptions have significant differences from each another)
- 4- Diverse (The descriptions are considerably different from each another)
- 3- Slightly Diverse (The descriptions tend to be similar but not the same)
- 2- Almost not Diverse (The descriptions are almost the same)
- 1- Not Diverse at all (The descriptions are identical or almost identical)

**Figure A.2:** Amazon mechanical turk example question for evaluating the quality and diversity of a set of referring expressions.