Easy Things First: Installments Improve Referring Expression Generation for Objects in Photographs

Sina Zarrieß **David Schlangen**
Dialogue Systems Group // CITEC // Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
first.last@uni-bielefeld.de

Abstract

Research on generating referring expressions has so far mostly focussed on "one-shot reference", where the aim is to generate a single, discriminating expression. In interactive settings, however, it is not uncommon for reference to be established in "installments", where referring information is offered piecewise until success has been confirmed. We show that this strategy can also be advantageous in technical systems that only have uncertain access to object attributes and categories. We train a recently introduced model of grounded word meaning on a data set of REs for objects in images and learn to predict semantically appropriate expressions. In a human evaluation, we observe that users are sensitive to inadequate object names - which unfortunately are not unlikely to be generated from low-level visual input. We propose a solution inspired from human task-oriented interaction and implement strategies for avoiding and repairing semantically inaccurate words. We enhance a word-based REG with context-aware, referential installments and find that they substantially improve the referential success of the system.

1 Introduction

A speaker who wants to refer to an object in a visual scene will try to produce a referring expression (RE) that (i) is semantically adequate, i.e. accurately describes the visual properties of the target referent, and (ii) is pragmatically and contextually appropriate, i.e. distinguishes the target from other objects in the scene but does not overload the listener with unnecessary information. Figure 1 illustrates this with two examples from a corpus of REs collected from human subjects for objects in images (Kazemzadeh et al., 2014).

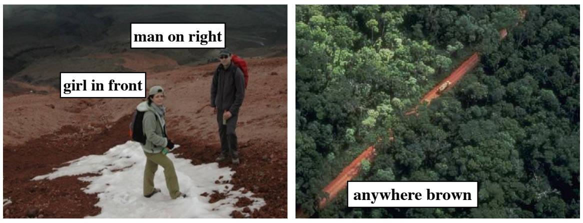


Figure 1: Example images and REs from the ReferIt corpus (Kazemzadeh et al., 2014)

Research on referring expression generation (REG) has mostly focussed on (ii), modeling pragmatic adequacy in attribute selection tasks, using as input a fully specified, symbolic representation of the visual attributes of an object and its distractors in a scene (Dale and Reiter, 1995; Krahmer and Van Deemter, 2012).

In this paper, we follow a more recent trend (Kazemzadeh et al., 2014; Gkatzia et al., 2015) and investigate REG on real-world images. In this setting, a low-level visual representation of an image (a scene) segmented into regions (objects), including the region of the target referent, constitutes the input. This task is closely related to the recently very active field of image-to-text generation, where deep learning approaches have been used to directly map low-level visual input to natural language sentences, e.g. (Vinyals et al., 2015; Chen and Lawrence Zitnick, 2015; Devlin et al., 2015). Similarly, we propose to cast REG on images as a word selection task. Thus, we base this work on a model of perceptually grounded word meaning, which associates words with classifiers that predict their semantic appropriateness given the low-level visual features of an object (Kennington and Schlangen, 2015). As our first contribution, we train this model on the ReferIt corpus (Kazemzadeh et al., 2014) and define decoding mechanisms tailored to REG.

Large-scale recognition of objects and their attributes in images is still a non-trivial task. Consequently, REG systems now face the challenge of dealing with semantically inadequate expressions. For instance, in Figure 1, the system might not precisely distinguish between *man* or *woman* and generate an inadequate, confusing RE like *man in the middle*. Therefore, we focus on evaluating our system in an object identification task with users, in contrast to previous approaches to REG on images (Mao et al., 2015). In order to assess possible sources of misunderstanding more precisely, our set-up also introduces a restricted form of interaction: instead of measuring "one-shot" performance only, users have three trials for identifying a referent. In this set-up, we find that different parameter settings of the systems (e.g. their visual inputs) have a clear effect on the referential success rates, while automatic evaluation measures reflect the interactive effectiveness rather poorly.

Research on reference in human interaction has noticed that conversation partners try to minimize their joint effort and often prefer to present simple expressions that can be expanded on or repaired, if necessary (Clark and Wilkes-Gibbs, 1986). This strategy, called "referring in installments" is very effective for achieving common ground in task-oriented interaction (Fang et al., 2014) and is attested in dialogue data (Striegnitz et al., 2012). The connection between reference in installments on the one and the status of distractors and distinguishing expressions on the other hand is relatively unexplored, though it seems natural to combine the two perspectives (DeVault et al., 2005). Figure 1 shows an example for very a simple but highly effective expression - it mentions color as a salient and distinguishing property while avoiding a potentially unclear object name.

As our second contribution, we extend our probabilistic word selection model to work in a simple interactive installment component that tries to avoid semantically inadequate words as much as possible and only expands the expression in case of misunderstanding. We present an algorithm that generates these installments depending on the context, based on ideas from traditional REG algorithms like (Dale and Reiter, 1995). We find that a context-aware installment strategy greatly improves referential success as it helps to avoid and repair misunderstandings and offers a combined treatment of semantic and pragmatic adequacy.

2 Background

2.1 Approaches to REG

"One-shot REG" Foundational work in REG has investigated attribute selection algorithms (Dale and Reiter, 1995) that compute a *distinguishing* referring expression for an object in a visual scene, which is defined as a target object r, set of distractor objects $D = \{d_1, d_2, d_3, ...\}$ and a set of attributes $A = \{type, position, size, color, ...\}$. A manually specified database typically associates the target and distractors in D with atomic values for each attribute, cf. (Krahmer and Van Deemter, 2012). In this setting, an attribute $a_1 \in A$ is said to *rule out* a distractor object from D, if the target and distractor have different values. This is mostly based on the assumption that we have objects of particular types (e.g. people, furniture, etc.) and that the system has perfect knowledge about these object types and, consequently, about potential distractors of the target. This does not apply to REG on real-world images which, as we will show in this paper, triggers some new challenges and research questions for this field. Subsequent work has shown that human speakers do not necessarily produce minimally distinguishing expressions (van Deemter et al., 2006; Viethen and Dale, 2008; Koolen et al., 2011), and has tried to account for the wide range of factors - such as different speakers, modalities, object categories - that are related to attribute selection, cf. (Mitchell et al., 2010; Koolen and Krahmer, 2010; Clarke et al., 2013; Tarenskeen et al., 2015).

Task-oriented REG has looked at reference as a collaborative process where a speaker and a listener try to reach a common goal (Clark and Wilkes-Gibbs, 1986; Heeman and Hirst, 1995; DeVault et al., 2005). Given the real-time constraints of situated interaction, a speaker often has to start uttering before she has found the optimal expression, but at the same time, she can tailor, extend, adapt, revise or correct her referring expressions in case the listener signals that he did not understand. Thus, human speakers can flexibly split and adapt their REs over several utterances during an interaction, a phenomenon called "reference in installments". In a corpus analysis of the S-GIVE domain, (Striegnitz et al., 2012) showed that installments are pervasive in human-human interaction in a task-oriented environment. However, while there has been research on goal-oriented and situated REG (Stoia et al., 2006; Kelleher and Kruijff, 2006; Striegnitz et al., 2011; Garoufi and Koller, 2013), installments have been rarely implemented and empirically tested in interactive systems. A noticeable exception is the work by Fang et al. (2014) who use reinforcement learning to induce an installment strategy that is targeted at robots that have uncertain knowledge about the objects in their environment. Using relatively simple computer-generated scenes and a standard representations of objects as sets of attributes, they learn a strategy that first guides the user to objects that the system can recognize with high confidence. Our work is targeted at more complex scenes in real-world images and large domains where no a priori knowledge about object types and their attributes is given. Mao et al. (2015) use a convolutional neural network and an LSTM to generate REs directly and on the same data sets as we do in this paper, but they only report automatic evaluation results.

2.2 The ReferIt corpus

We train and evaluate our system on the ReferIt data set collected by Kazemzadeh et al. (2014). The basis of the corpus is a collection of "20,000 still natural images taken from locations around the world" (Grubinger et al., 2006), which was augmented by Escalante et al. (2010) with segmentation masks identifying objects in the images (an average of 5 objects per image). This dataset also provides manual annotations of region labels and a vector of visual features for each region (e.g. region area, width, height, and color-related features). There are 256 types of objects (i.e. labels), out of which 140 labels are used for more than 50 regions (Escalante et al., 2010). Kazemzadeh et al. (2014) collected a large number of expressions referring to objects (for which segmentations exist) from these images (130k REs for 96k objects), using a game-based crowd-sourcing approach, and they have assembled an annotated test set.

2.3 The WAC model

Given a corpus of REs aligned with objects in images, we can train a model that predicts semantically appropriate words given the visual representation of an image region. We adopt the WAC ("words-as-classifiers") model (Kennington and Schlangen, 2015), which was originally used for reference resolution in situated dialogue. However, WAC is essentially a task-independent approach to predicting semantic appropriateness of words in visual contexts and can be flexibly combined with task-dependent decoding procedures.

The WAC model pairs each word w in its vocabulary V with an individual classifier that maps the low-level, real-valued visual properties of an object o to a semantic appropriateness score. In order to learn the meaning of e.g. the word *red*, the visual properties of all objects described as *red* in a corpus of REs are given as positive instances to a supervised (logistic regression) learner. Negative instances are randomly samples from the complementary set of utterances (e.g. not containing *red*).

We used this relatively simple model in our work, because first of all we wanted to test wether it scales from a controlled domain of typical reference game scenes (Kennington and Schlangen, 2015) to real-world images. Second, as compared to standard object recognisers that predict abstract image labels annotated in e.g. ImageNet (Deng et al., 2009), this model directly captures the relation between actual words used in REs and visual properties of the corresponding referents. Following (Schlangen et al., 2016), we can easily base our classifiers on such a high-performance convolutional neural network (Szegedy et al., 2015), by applying it on our images and extracting the final fully-connected layer before the classification layer (see Section 3.1).

3 Word-based REG for Image Objects

We describe a word selection model for REG on images, which reverses the decoding procedure of our reference resolution model (Kennington and Schlangen, 2015; Schlangen et al., 2016). The main question we pursue here is whether we can predict semantically adequate words for visually represented target objects in real-world images and achieve communicative success in a task-oriented evaluation.

3.1 A Basic Algorithm for REG with WAC

Given a visual representation of an object, we can apply all word classifiers from the vocabulary of our WAC model and obtain an appropriateness ranking over words. As these WAC scores do not reflect appropriateness in the linguistic context, i.e. the previously generated words, we combine them with simple language model (bigram) probabilities (LM) computed on our corpus. The combination of WAC and LM scores is used to rank our vocabulary with respect to appropriateness given the visual features of the target referent and linguistic context.

Algorithm 1 shows our implementation of the decoding step, a beam search that iteratively adds n words with the highest combined LM and WAC score to a its agenda and terminates after a prespecified number of maximum steps.

The algorithm takes the number of iterations as input, so it searches for the optimal RE given a fixed length. Deciding how many words have to be generated is very related to deciding how many attributes to include in more traditional REG. As a first approach, we have trained an additional regression classifier that predicts the length of the RE, given the number of objects in the scene and the visual properties of the target.

Algorithm 1 Baseline REG with WAC

```
function WORD-GEN($object, maxsteps, V$)
    $Agenda \leftarrow \{`start'\}$
    for $n \in 0..maxsteps$ do
        $Beam \leftarrow \emptyset$
        for $re \in Agenda$ do
            $w_{-1} \leftarrow$ LAST($re$)
            for $w \in$ BIGRAMS($w_{-1}, V$) do
                $s =$ WAC($w, object$) + LM($w, w_{-1}$)
                $re_{new} \leftarrow$ APPEND($re, word$)
                $Beam \leftarrow Beam \cup \{(re_{new}, s)\}$
            end for
        end for
        $Agenda \leftarrow$ K-BEST($Beam, k$)
    end for
    return K-BEST($Agenda, 1$)
end function
```

3.2 Experimental Set-up

Data We use the same test set as Kazemzadeh et al. (2014) that is divided into the 3 subsets, each containing 500 objects: "**Set A** contains objects randomly sampled from the entire dataset, **Set B** was sampled from the most frequently occurring object categories in the dataset, excluding the less interesting categories, **Set C** contains objects sampled from images that contain at least 2 objects of the same category, excluding the less interesting categories."[1] For each object, there are 3 human-generated reference REs. We train the WAC model on the set of images that are not contained in the test set, which amounts to 100384 REs.

The classifiers We use Schlangen et al. (2016)'s WAC model that is a trained on the REFERIT data (Kazemzadeh et al., 2014) based on the SAIAPR collection (Grubinger et al., 2006). We train binary logistic regression classifiers (with $\ell 1$ regularisation) for the 400 most frequent words from the training set.[2] During training, we only consider non-relational expressions, as words from relational expressions would introduce further noise. Each classifier is trained with the same balance of positive and negative examples, a fixed ratio of 1 positive to 7 negative. Additionally, we train a regression classifier that predicts the expected length of the RE given the visual features of the target object and the number of objects in the entire scene. We also train a simple bigram language model on the data.

Feature sets In this experiment, we manipulate the features sets of the underlying word classifiers. We train it on (i) a small set of 27 low-level visual features extracted and provided by Escalante et al. (2010), called SAIAPR features below, and (ii) a larger set of features automatically learned by a state-of-the-art convolutional neural network, "GoogLeNet" (Szegedy et al., 2015). We derive representations of our visual inputs with this CNN, that was trained on data from the ImageNet corpus (Deng et al., 2009), and extract the final fully-connected layer before the classification layer, to give us a 1024 dimensional representation of the region. We augment this with 7 features that encode information about the region relative to the image: the (relative) coordinates of two corners, its (relative) area, distance to the center, and orientation of the image. The full representation hence is a vector of 1031 features. The feature extraction for (ii) is described in more detail in (Schlangen et al., 2016). Generally, the SAIAPR features represent interpretable visual information on position, area, and color of an image region, they could be associated with particular visual attributes. This is not possible with the GoogLeNet features.

[1] Where objects mostly located in the background like 'sky', 'mountain' are considered to be less interesting.

[2] We used `scikit learn` (Pedregosa et al., 2011).

3.3 Automatic Evaluation

To the best of our knowledge, end-to-end REG performance has not been reported on the ReferIt data set before. Table 1 shows corpus-based BLEU and NIST measures calculated on the test set (using 3 references for each RE). The results indicate a minor gain of the GoogLeNet features. We also evaluate a version of the GoogLeNet-based system that instantiates the beam search with the gold length of the RE from the corpus (GoogLeNet$_{glen}$). This leads to a small improvement in BLEU and NIST, indicating that the length prediction is not a critical factor.

	BLEU		NIST	
	1-gram	2-gram	1-gram	2-gram
SAIAPR	0.33	0.19	1.5	1.7
GoogLeNet	0.35	0.21	1.9	2.3
GoogLeNet$_{glen}$	0.38	0.19	2.0	2.6

Table 1: Automatic evaluation for word-based REG systems

3.4 A Game-based Human Evaluation

Set-up In parallel to the reference game in (Kazemzadeh et al., 2014), we set up a game between a computer that generates REs and a human player who clicks on the location of the described object that he identifies based on the RE. After each click, the GUI presents some canned feedback and informs the player whether he clicked on the intended object. In case of an unsuccessful click, the player has two more trials. In the following, we report the success rates with respect to each trial and the different test sets. This set-up will trigger a certain amount of user guesses such that the success rates do not correspond perfectly to semantic accuracies. But it accounts for the increased difficulty as well as the interactive nature of the task. See Section 4.4 for an analysis of learning effects in this set-up and (Gatt et al., 2009; Belz and Hastie, 2014) for general discussion on REG and NLG evaluation.

	Success rate/ trial			Error
	1st	2nd	3rd	red.
SAIAPR	32.2	40.3	46.3	20.8
GoogLeNet	41.6	53.4	59.1	29.9
GoogLeNet$_{glen}$	37.6	51	58.7	33.8
human	90.6	94.6	98.3	81.9

Table 2: Human success and error reduction rates in object identification task, for different sets of visual features

For each player, we randomly sampled the games from the entire test set, but balanced the items so that they were equally distributed across the 3 test subsets A, B, C (see above) and the three systems. We also included human REs from the corpus. In total, we collected 1201 games played by 8 participants.

Results In Table 2, we report the cumulative success rates for the different systems across the different trials, i.e. the success rate in the 3rd trial corresponds to the overall proportion of successfully identified referents. First of all, this suggests that the differences in performance between the systems is much bigger in terms of their communicative effectiveness as in terms of the corpus-based measures (Table 1). Thus, on the one hand, the GoogLeNet features are clearly superior to SAIAPR, whereas differences between GoogLeNet and GoogLeNet$_{glen}$ are minor. Interestingly, the GoogLeNet features improve 1st trial as well as overall success, leading to a much better error reduction rate[3] in object identification between the first and third trial. This means that, here, humans are more likely to recover from misunderstandings and indicates that REs generated by the SAIAPR system are more semantically inadequate.

	Success rate (3rd trial)		
	Set A	Set B	Set C
SAIAPR	35.7	63.8	40.7
GoogLeNet	57	67.7	53.1
GoogLeNet$_{glen}$	50	74	53
human	99.1	99	96.5

Table 3: Human success rates for baseline REG systems trained on different visual feature sets

In Table 3, we report the overall success rates for the different test sets. All systems have a clearly higher performance on the B Set which contains the most frequent object types. Surprisingly, all systems have a largely comparable performance on Set A and C whereas only C contains images with distractors in the sense of traditional REG. This shows that describing objects which belong to an infrequent type in a semantically adequate way, which is necessary in Set A, is equally challenging as reaching pragmatic adequacy which is called for in Set C.

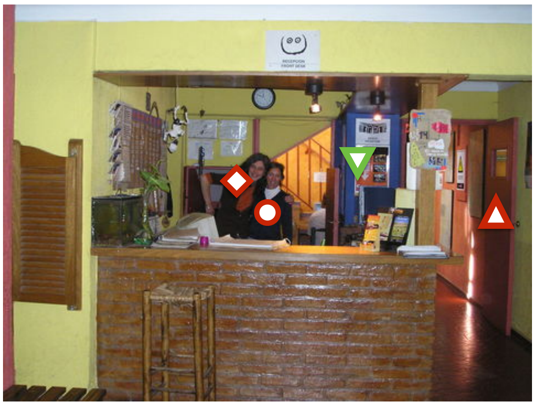
[3] Calculated as $(\#error_{1st} - \#error_{3rd})/\#error_{1st}$

3.5 Error Analysis

When users click on a distractor object instead of the intended referent in our object identification task, there can be several sources of misunderstanding. For instance, it is possible that the system produced REs that are interpretable but not specific and distinguishing enough to identify the target. It is also possible that the system selected words that are semantically inadequate such that the expression becomes completely misleading. We can get some insight into possible sources of misunderstanding by comparing the the clicked-on distractor objects to their intended target, using the object labels annotated for each image region (see Section 2.2).

The analysis of mismatches between the expected label of the target and the label of the object actually clicked on by the user reveals that many errors are due to semantic inadequacies and apparently severe interpretation failures: Looking at the total number of clicks on distractor objects, 80% are clicks on a distractor with a different label than the target.[4], e.g. the user clicked on a 'tree' instead of a 'man'. This is clear evidence for semantic inadequacies, suggesting that the systems often generate an inadequate noun for the object type. An example for such a label mismatch is shown in Figure 2 where the system generated "person" for referring to a "sign", such that the user first clicked on distractor objects that are persons.

Similarly, we can get some evidence about how users try to repair misunderstandings, by comparing a distractor clicked on in the first trial to another distractor clicked on in the subsequent second, or third trial. Interestingly, we find that users do not seem to be aware of the fact that the system does not always generate the correct noun and do not generally try to click on objects with a different label. Only in 39% of the unsuccessful second trials, users decided for a distractor object with a different label, even though the first click had been unsuccessful. For instance, in Figure 2, the user clicked on the other person in the image in the second trial, although this referent is clearly not on the right. This suggests that users do not easily revise their RE interpretation with respect to the intended type of referent.



re$_{generated}$: "person on the right"
re$_{human}$: "sign on the blue shelf in the back"

Figure 2: Example for an unsuccessful trial in object identification; first click: ○, second click: ◇, third click: △, target: ▽)

Moreover, we can compare the different distractor clicks with respect to their spatial distance to the target. We find that after an unsuccessful first trial, users click on an object that has a greater distance to the target in 70% of the cases (as e.g. in Figure 2). This means that users often try to repair the misunderstanding with respect to the intended location, rather than with respect to the intended object type. Intuitively, this behaviour makes sense: a human speaker is more likely to confuse e.g. *left* and *right* than e.g. *man* and *tree*. From the perspective of the system this is a problematic situation: words like *left* and *right* are much easier to generate (based on simple positional features) than nouns like *man* and *tree*.

4 Towards interactive, contextual REG

In this Section, we extend our word-based REG to deal with semantic inadequacies. We take a first step towards interactive REG and implement installments, a pervasive strategy in human task-oriented interaction. The main idea is that the system should try to avoid semantically inadequate expressions wherever possible and, if misunderstanding occurs, try to react appropriately.

4.1 Procedure

When a speaker or system *refers in installments*, they do not need to generate an RE in one shot, but can start with an initial, simple RE that is extended or reformulated if this becomes necessary in the interaction, i.e. if the listener cannot identify the referent. This setting is a straightforward extension of our game-based evaluation in Section 3.4, where users had 3 trials for identifying a referent: instead of generating a single RE for the target

[4] The percentage varies between saiapr (86%), GoogLeNet (71%)

and presenting it in every trial, we now produce a triple (re_1, re_2, re_3), where re_1 will be used in the first trial, re_2 in the second trial, etc.

In this set-up, we want to investigate whether installments and reformulations help to avoid semantic inadequacies and improve referential success, i.e. whether a dynamic approach to REG compares favourably to the non-dynamic version of our system (see Section 3). This question is, however, closely linked to another, more intricate question: what is the best strategy to realize installments that, on the one hand, provide enough information so that a user can eventually identify the referent and, on the other hand, avoid misleading words? To date, even highly interactive systems do not generally treat installments, or if they do, only realise them via templates, e.g. (Stoia et al., 2006; Staudte et al., 2012; Garoufi and Koller, 2013; Dethlefs and Cuayáhuitl, 2015). As pointed out by Liu et al. (2012), data-driven approaches are not straightforward to set-up, due to the "mismatched perceptual basis" between a human listener and an REG system.

Based on the insights of our error analysis in Section 3.4, we will rely on a general installment strategy that is mostly targeted at avoiding semantically inadequate object names, and emphasizing the fact that location words generated by the system convey more reliable information. We have implemented two versions of this general strategy: (i) pattern-based installments that always avoid object names in their initial expression and dynamically extend this if necessary, (ii) context-dependent installments that condition the initial expression on the complexity of the scene and extend the initial expression accordingly, inspired by standard approaches to attribute selection in REG (Krahmer and Van Deemter, 2012). Thus, we do not test initial or reformulated expressions in isolation, but the strategy as a whole, which is similar to (Fang et al., 2014).

4.2 Pattern-based Installments

This system generates a triple of REs for each image, corresponding to the respective trials in the object identification task. The triple for pattern-based installments is defined as follows:

- re_1: a short RE that only contains location words, e.g. *bottom left*
- re_2: a longer RE that contains location words and an object name, e.g. *the car on the left*
- re_3: a reformulation of re_2 that hedges the object name and suggests an alternative object name, e.g. *vehicle or bottle on the left*

Figure 3(a) illustrates a case where this pattern is useful: the target is a horse, the biggest and most salient object in the image, which can be easily identified with a simple locative expression. As horses are not frequent in the training data, the system unfortunately generates *hat guy* as the most likely object name. This RE would be very misleading indeed if presented to a listener, as one of the distractors actually is a person with a hat.

Generation Procedure In order to generate the above installment triples with our REG system, we simply restrict the vocabulary of the underlying WAC-model. Thus, we divided the 400 word classifiers into the following subsets:

- V_1: 20 location words (manually defined)
- V_2: V_1 + 183 object names (extracted from annotated section of the ReferIt corpus)
- V_3: entire vocabulary

This basic installment-based system does not use V_3 (but see below). For generating the hedge of the object name in the third trial (re_3) we use the top-second and top-third word from the ranking that WAC produces over all object type words given the visual features of the target.

4.3 Context-dependent Installments

Our context-dependent installment strategy determines the initial RE (re_1) based on the surrounding scene and generates subsequent reformulations (re_2,re_3) accordingly.

Initial REs and Distractors As we do not have a symbolic representation of the distractor objects and their properties, we use the word-based REG system to decide whether an RE can be expected to be distinguishing for the target in the scene. This is similar to (Roy, 2002). Algorithm 2 shows the procedure for determining the initial RE (re_1). Same as before, we restrict the vocabulary of the underlying WAC model, e.g. to contain only location words. But now, we apply the word generation function to the target object and to all the other objects in the set of distractors (D). If the algorithm generates an identical chunk for the target and one of its distractors, it continues with a less restricted vocabulary and a longer expression. It terminates when it has found an RE that is optimal only for the target. This algorithm proceeds on the level of chunks, instead of single words, as e.g. location is often described by several words (e.g. *bottom left*).

Algorithm 2 A Context-aware REG Algorithm

```
function INC-GEN($object, maxsteps, D, V$)
    for $n \in 2..maxsteps$ do
        $V_n \leftarrow$ RESTRICT($V, n$)
        $re \leftarrow$ WORD-GEN($object, V_n$)
        for $d \in D$ do
            $re_d \leftarrow$ WORD-GEN($d, V_n$)
            if $re_d = re$ then
                break
            end if
        end for
        return $re$
    end for
end function
```

As we found that the linguistic quality degrades for longer REs, we limit the maximal RE length to 6 words. We obtain 3 types of initial REs predicted to be distinguishing for a target by Algorithm 2:

- ref_{loc}: 2 word RE, only location words (V_1), Figure 3(a)
- ref_{object}: 4 word RE, location words and object names (V_2), Figure 3(b)
- ref_{att}: 6 word RE, all attributes from the entire vocabulary (V_3), Figure 3(c)

On our test set, this produces distinguishing REs for all targets, except 4 cases for which we use an initial 6 word RE as well.

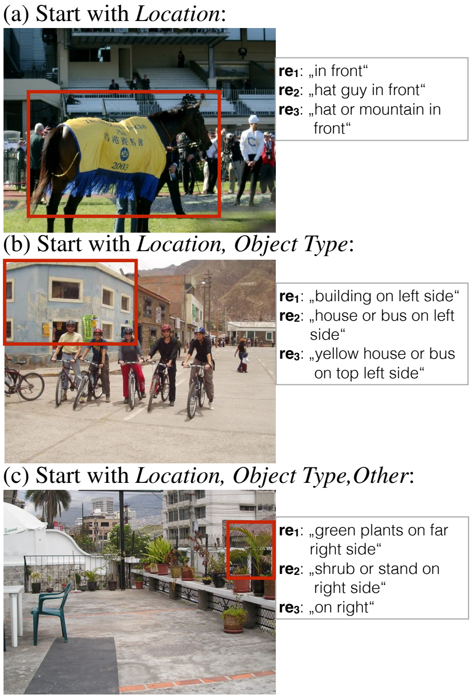
Reformulations We have several options for generating the reformulation REs (re_2,re_3) - e.g. hedging the object name, extending the RE with more words, removing potentially misleading words, etc. - which are more or less appropriate, depending on the initial RE predicted by Algorithm 2. Therefore, we implemented the following types of installment triples that dynamically extend or reduce the initial RE:

1. (ref_{loc}, ref_{object}, $ref_{object,hedge}$), this corresponds to the pattern in Section 4.2
2. (ref_{object}, $ref_{object,hedge}$,ref_{att})
3. (ref_{att}, $ref_{att,hedge}$,ref_{loc})

Figure 3 shows examples for each triple.

4.4 Human Evaluation

Set-up We use the task-oriented setup from Section 3.4 with 3 trials per image. But instead of presenting the same RE in each trial, the system now updates the phrases according to the RE triples described above. We have recruited 5 players and collected 1200 games, split equally between the pattern-based installment (Section 4.2) and the context-dependent installment strategy (Section 4.3). In this evaluation, we only use word classifiers trained on GoogLeNet features.



(a) Start with *Location*:

re1: „in front“
re2: „hat guy in front“
re3: „hat or mountain in front“

(b) Start with *Location, Object Type*:

re1: „building on left side“
re2: „house or bus on left side“
re3: „yellow house or bus on top left side“

(c) Start with *Location, Object Type,Other*:

re1: „green plants on far right side“
re2: „shrub or stand on right side“
re3: „on right“

Figure 3: Examples for context-dependent installments

Results Table 4 shows that even the simple, pattern-based installment system improves the 1st trial success rate compared to the non-interactive baseline (the GoogLeNet-based system from Section 3) and is clearly superior with respect to its overall success and error reduction rate over trials. This suggests that a fair amount of target objects can be identified by users based on very simple, locative REs as semantically inadequate object names are avoided. Another important finding here is the high rate of error reduction during the 2nd and 3rd trial achieved by the installment-based system. In the non-interactive system, users did not have additional cues for repairing their misunderstanding and probably guessed other possible targets in individual, more or less systematic ways. Apparently, even simple strategies for extending and hedging the initially presented RE provide very helpful cues for repairing initial misunderstandings.

As we expected, the pattern-based installment system is clearly improved by our context-dependent approach to generating installments. This systems seems to strike a much better balance between generating simple expressions that avoid

	Success rate/ trial			Error
	1st	2nd	3rd	red.
No install.	41.6	53.4	59.1	29.9
Pattern install.	46.8	69.2	80.9	64.1
Contextual install.	50.5	74.9	86	71.71

Table 4: Human evaluation for installment-based REG systems

inadequate object names on the one and contextually appropriate expressions on the other hand. It improves the pattern-based installments in terms of 1st trial success rate and overall success and error reduction rate.

The finding that installment strategies should be combined with insights from traditional distractor-oriented REG is further corroborated when we compare the success rates on the different subsets of our test set, see Table 5. Thus, the performance of the context-dependent installment system is much more stable on the different subsets than the pattern-based system which has a clear dip in success rate on Set C, which contains target referents with distractors of the same object type. This result suggests that our approach to determine distinguishing REs based purely on predictions of word-based REG (Section 4.3) presents a viable solution for REG on images, where information on distractors is not directly assessable in the low-level representation of the scene.

	Success rate (3rd trial)		
	Set A	Set B	Set C
No install.	57	67.7	53.1
Pattern install.	80.8	84.3	77.5
Contextual install.	86	87.5	84.5

Table 5: Human evaluation on different test sets for installment-based REG systems

Finally, the graph in Figure 4 shows the average success rates over time and provides more evidence for the effectiveness of installments. We observe a clear learning effect in the non-interactive system, meaning that users faced unexpected interpretation problems due to inaccurate expressions, but adapted to the situation to some extent. In contrast, both installment systems have stable performance over time, which indicates that system behaviour is immediately understandable and predictable for human users.

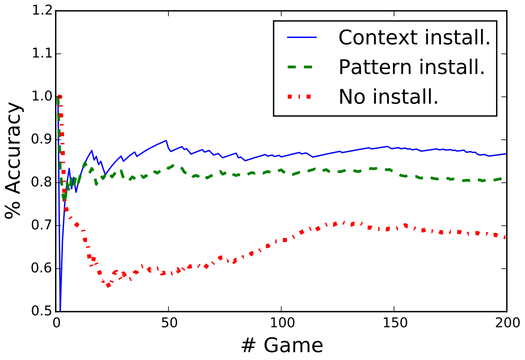


Figure 4: Participants' success rates in object identification over time

5 Discussion and Conclusion

We have presented an REG system that approaches the task as a word selection problem and circumvents manual specification of attributes in symbolic scene representations as required in traditional REG (Krahmer and Van Deemter, 2012), or manual specification of attribute-specific functions that map particular low-level visual features to attributes or words as in (Roy, 2002; Kazemzadeh et al., 2014). This knowledge-lean approach allows us to use automatically learned ConvNet features and obtain a promising baseline that predicts semantically appropriate words based on visual object representations.

We have argued and demonstrated that REG in more realistic settings greatly benefits from a task-oriented, interactive account and should explore principled strategies for repairing and avoiding misunderstandings due to semantically inaccurate REs. In order to achieves this, we have augmented our approach with some manually designed installment strategies. An obvious direction for future work is to automatically induce such a strategy, based on confidence measures that automatically predict the trust-worthiness of a word for an object.

Another extension that we have planned for future work is to implement relational expressions, similar to (Kennington and Schlangen, 2015). Based on relational expressions, we will be able to generate reformulations and installments tailored to the interaction with the user. For instance, a very natural option for installments is to relate the wrong target object clicked on by the user to the intended target, e.g. something like *to the left of that one, the bigger object*.

Acknowledgments

We acknowledge support by the Cluster of Excellence "Cognitive Interaction Technology" (CITEC; EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG).

References

Anja Belz and Helen Hastie. 2014. Comparative evaluation and shared tasks for nlg in interactive systems. In Amanda Stent and Srinivas Bangalore, editors, *Natural Language Generation in Interactive Systems*, pages 302–350. Cambridge University Press. Cambridge Books Online.

Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2422–2431.

Herbert H Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22(1):1–39.

Alasdair DF Clarke, Micha Elsner, and Hannah Rohde. 2013. Where's wally: the influence of visual salience on referring expression generation. *Frontiers in psychology*, 4.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.

Jia Deng, W. Dong, Richard Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Nina Dethlefs and Heriberto Cuayáhuitl. 2015. Hierarchical reinforcement learning for situated natural language generation. *Natural Language Engineering*, 21(03):391–435.

David DeVault, Natalia Kariaeva, Anubha Kothari, Iris Oved, and Matthew Stone. 2005. An information-state approach to collaborative reference. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 1–4.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 100–105, Beijing, China, July. Association for Computational Linguistics.

Hugo Jair Escalante, Carlos a. Hernández, Jesus a. Gonzalez, a. López-López, Manuel Montes, Eduardo F. Morales, L. Enrique Sucar, Luis Villaseñor, and Michael Grubinger. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 114(4):419–428.

Rui Fang, Malcolm Doering, and Joyce Y. Chai. 2014. Collaborative Models for Referring Expression Generation in Situated Dialogue. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Konstantina Garoufi and Alexander Koller. 2013. Generation of effective referring expressions in situated context. *Language and Cognitive Processes*.

Albert Gatt, Anja Belz, and Eric Kow. 2009. The tuna-reg challenge 2009: Overview and evaluation results. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 174–182. Association for Computational Linguistics.

Dimitra Gkatzia, Verena Rieser, Phil Bartie, and William Mackaness. 2015. From the virtual to the real world: Referring to objects in real-world spatial scenes. In *Proceedings of EMNLP 2015*. Association for Computational Linguistics.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 benchmark: a new evaluation resource for visual information systems. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2006)*, pages 13–23, Genoa, Italy.

Peter A Heeman and Graeme Hirst. 1995. Collaborating on referring expressions. *Computational linguistics*, 21(3):351–382.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 787–798, Doha, Qatar.

John D Kelleher and Geert-Jan M Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 1041–1048.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China, July. Association for Computational Linguistics.

Ruud Koolen and Emiel Krahmer. 2010. The d-tuna corpus: A dutch dataset for the evaluation of referring expression generation algorithms. In *LREC*.

Ruud Koolen, Albert Gatt, Martijn Goudbeek, and Emiel Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Changsong Liu, Rui Fang, and Joyce Y Chai. 2012. Towards mediating shared perceptual basis in situated dialogue. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 140–149. Association for Computational Linguistics.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2015. Generation and comprehension of unambiguous object descriptions. *CoRR*, abs/1511.02283.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2010. Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference*, pages 95–104. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Deb K Roy. 2002. Learning visually grounded words and syntax for a scene description task. *Computer Speech & Language*, 16(3):353–385.

David Schlangen, Sina Zarriess, and Casey Kennington. 2016. Resolving references to objects in photographs using the words-as-classifiers model. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Maria Staudte, Alexander Koller, Konstantina Garoufi, and Matthew W Crocker. 2012. Using listener gaze to augment speech generation in a virtual 3d environment. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society. To appear*.

Laura Stoia, Darla Magdalene Shockley, Donna K Byron, and Eric Fosler-Lussier. 2006. Noun phrase generation for situated dialogs. In *Proceedings of the fourth international natural language generation conference*, pages 81–88. Association for Computational Linguistics.

Kristina Striegnitz, Alexandre Denis, Andrew Gargett, Konstantina Garoufi, Alexander Koller, and Mariët Theune. 2011. Report on the second second challenge on generating instructions in virtual environments (give-2.5). In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 270–279.

Kristina Striegnitz, Hendrik Buschmeier, and Stefan Kopp. 2012. Referring in installments: a corpus study of spoken object references in an interactive virtual environment. In *Proceedings of the Seventh International Natural Language Generation Conference*, pages 12–16.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR 2015*, Boston, MA, USA, June.

Sammie Tarenskeen, Mirjam Broersma, and Bart Geurts. 2015. hand me the yellow stapler or hand me the yellow dress: Colour overspecification depends on object category. page 140.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the Fourth International Natural Language Generation Conference*, pages 130–132. Association for Computational Linguistics.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.