From the Virtual to the Real World: Referring to Objects in Real-World Spatial Scenes

Dimitra Gkatzia, Verena Rieser
Department of Computer Science
Heriot-Watt University
Edinburgh EH14 4AS, UK
{d.gkatzia,v.t.rieser}@hw.ac.uk

Phil Bartie
School of Natural Sciences
University of Stirling
Stirling FK9 4LA, UK
phil.bartie@stir.ac.uk

William Mackaness
School of GeoSciences
University of Edinburgh
Edinburgh EH8 9XP, UK
william.mackaness@ed.ac.uk

Abstract

Predicting the success of referring expressions (RE) is vital for real-world applications such as navigation systems. Traditionally, research has focused on studying Referring Expression Generation (REG) in virtual, controlled environments. In this paper, we describe a novel study of spatial references from real scenes rather than virtual. First, we investigate how humans describe objects in open, uncontrolled scenarios and compare our findings to those reported in virtual environments. We show that REs in real-world scenarios differ significantly to those in virtual worlds. Second, we propose a novel approach to quantifying image complexity when complete annotations are not present (e.g. due to poor object recognition capabilities), and third, we present a model for success prediction of REs for objects in real scenes. Finally, we discuss implications for Natural Language Generation (NLG) systems and future directions.

1 Introduction

REG has attracted considerable interest in the NLG community over the past 20 years (Krahmer and van Deemter, 2011; Gatt et al., 2014). While initially, the standard evaluation metric for REG was human-likeness, as compared to human corpora similarity as in TUNA (Gatt et al., 2009), the field has moved on to evaluating REG effectiveness by measuring task success in virtual interactive environments (Byron et al., 2009; Gargett et al., 2010; Janarthanam et al., 2012). Virtual environments however eliminate real-world uncertainty, such object recognition errors or cluttered scenes. In this paper, we investigate whether the lessons learnt in virtual environments can be transferred to real-world scenes. We consider the case where we are uncertain about the scene itself, i.e. we assume that the complexity of the scene is hidden and we are interested in identifying a specific object, and thus our work differs from approaches that generate descriptions for images such as (Mitchell et al., 2012; Yang et al., 2011; Yatskar et al., 2014).

Related work has focused on computer generated objects (van Deemter et al., 2006; Viethen and Dale, 2008), crafts (Mitchell et al., 2010), or small objects in a simple background (Mitchell et al., 2013a; FitzGerald et al., 2013). One notable exception is the recent work by Kazemzadeh et al. (2014), who investigate referring expressions of objects in "complex photographs of real-world cluttered scenes". They report that REs are heavily influenced by the object type. Here, we are interested in studying REs for visual objects in urban scenes. As the success of a RE is heavily dependent on the complexity of the scene as well as its linguistic features, we are interested in modelling and thus predicting the success of a RE.

Initially, this paper presents and analyses a novel, real-world corpus REAL (to be released) – “Referring Expression Anchored Language” (Section 2), and compares the findings to those reported in virtual worlds (Gargett et al., 2010). We then provide a detailed analysis of how syntactic and semantic features contribute to the success of REs (Sections 4.1, 4.2, 4.3), accounting for unobservable latent variables, such as the complexity of the visual scene (as described in Section 3). Finally, we summarise our work and discuss the implications of our work for NLG systems (Section 5). The dataset and models will be released.

2 The REAL Corpus

The REAL corpus contains a collection of images of real-world urban scenes (Fig. 1) together with verbal descriptions of target objects (see Fig. 2)
generated by humans, paired with data on how successful other people were able to identify the same object based on these descriptions (Fig. 3). The data was collected through a web-based interface. The images were taken in Edinburgh (Scotland, UK), very early one summer morning. This was necessary to reduce the occlusion of city objects from buses and crowds, and to minimise lighting and weather variations between images.

2.1 Experimental Setup

There were 190 participants recruited (age between 16 to 71). Each participant was presented with an urban image (Fig. 1), where the target object was outlined by a yellow box (Fig. 2), and was asked to describe the target using free text. After completing a (self-specified) number of tasks, participants were then asked to validate descriptions provided by other participants by clicking on the object using previously unseen images (Fig. 3). Overall, 868 descriptions across 32 images were collected, averaging around 27 descriptions per image. The balance of generation and validations was adjusted to ensure that all descriptions were identified by at least 3 other participants, generating 2618 image tag verifications. All cases were manually checked to determine if the ‘correct’ (green) or ‘incorrect’ (red) target had been identified Fig. 3. Overall, 76.2% of human descriptions provided were successfully identified. For the experiments reported in following sections, we summarised answers categorised as ‘incorrect’, ‘ambiguous’ and ‘not found’ as unsuccessful.

2.2 Comparison to GIVE-2 Corpus

We now compare this data with human data generated for the GIVE-2 challenge (Gargett et al., 2010). In GIVE-2, the target objects have distinct attributes, such as colour and position. For instance, an effective RE in GIVE-2 could be “the third button from the second row”. In real-world situations though, object properties are less well defined, making a finite set of pre-defined qualities unfeasible. Consider, for instance, the building highlighted in Figure 2, for which the following descriptions were collected:

1. The Austrian looking white house with the dark wooden beams at the water side.
2. The white building with the x-shape balconies. It seems it’s new.
3. The white building with the balconies by the river.
4. Apartments with balconies.
5. The nearest house on right side. It’s black and white.
6. The white and black building on the far right, it has lots of triangles in its design.
7. The rightmost house with white walls and wood finishings.

It is evident that the REAL users refer to a variety of object qualities. We observe that all participants refer to the colour of the building (white, black and white, greyish-whitish) and some mention location (by the river, at the water side).

Experimental Factors influencing Task Performance: In REAL, task success is defined as the ability to correctly identify an object, whereas in GIVE-2, task success refers to the successful completion of the navigation task. In contrast to GIVE-2, not all REAL participants were able to correctly identify the referred objects (76.2% task
success). We assume that this is because GIVE-2 was an interactive setup, where the participants were able to engage in a clarification dialogue. **Gender:** In REAL, gender was not a significant factor with respect to task success (Mann-Whitney U test, \( p = 0.2 \)). **Length of REs (no. words):** In REAL, females tend to provide lengthier REs than males, however the difference is not statistically significant (Mann-Whitney U test, \( p = 0.58 \)). In GIVE-2, only German females produced significantly longer descriptions than their male counterparts. **Relation between length (no. words) and task success:** The REAL data shows a positive relationship between length and success rate, i.e. for a one word increase in length, the odds of correct object identification is significantly increased (\( p < 0.05 \), Logit), i.e. longer and more complex sentences lead to more successful REs.

### 3 Quantifying the Image Complexity

We assume that the complexity of the urban scene represented in the image is hidden due to the lack of semantic annotations. Our dataset does not include any quantifiable image descriptions, such as computer vision output as in (Mitchell et al., 2012) or manual annotations as in (Yatskar et al., 2014). In addition, the same RE might not always result in successful identification of an object due to scene complexity. In order to marginalise the effect of the scene complexity, we exploit the multiple available data points per image. This allows us to estimate the average success rate of each referring expression \( \overline{SR}_{RE} \) (the proportion of successful validations) and the average success rate of each image \( \overline{SR}_i \) (the proportion of the correctly identified objects in the image). We use \( \overline{SR}_i \) to marginalise over the (hidden) image complexity, where we assume that some pictures are inherently more complex than others and thus achieve lower success rates. Similar normalisation methods are used for user ratings to account for the fact that some users are more “tolerant” and in general give higher ratings (Jin and Si, 2004). We employ Gaussian normalisation (Resnick et al., 1994) to normalise image success rates by considering the following factors:

1. **Shift of average success rate per image:** some images are inherently easier than others and gain higher success rates, independently of the REs used. This factor can be accounted by subtracting average success rates of all images from the average rating of a specific image \( x \).

2. **Different ratings:** there are 27 REs per image on average, some of which are harder to understand than others, thus they gain lower success rates. To account for this, the success rates of each image are divided by the overall \( SR \) variance.

The normalised image success rate \( (NSR_i) \) per image \( x \) is defined by the following equation:

\[
NSR_i(x) = \frac{SR_i(x) - \overline{SR}_i}{\sqrt{\sum_n (SR_n(x) - \overline{SR}_n)^2}}
\]  

Using the \( (NSR_i) \), we now investigate the REs in terms of their linguistic properties, including automatically annotated syntactic features and manually annotated semantic features.

### 4 Modelling REG Success

Unlike previous work, we use both successful and unsuccessful REs in order to build a model that is able to predict the success or the failure of a RE.

#### 4.1 Syntactic Analysis of REG Success

We use the Stanford CoreNLP tool (Manning et al., 2014) to syntactically annotate the REs and we investigate which linguistic features contribute to the RE success in relation to the image complexity. Note that these analyses are based on normalised values, as discussed in Section 3).

**Predicting RE Success Rate \( (\overline{SR}_{RE}) \):** Initially, we compare successful and unsuccessful REs by taking the upper and lower quartiles and extracting their syntactic features, i.e. the top and bottom 25% of REs with respect to their average success rate, and group them into two groups. We then extract syntactic features of these two groups and compare their frequencies (occurrence per RE), means, and standard deviations (Table 3), and compare them using a t-test (\( p < 0.05 \)). The difference between successful and unsuccessful ex-
Table 3: Statistics regarding the linguistic features in successful vs unsuccessful referring expressions. (* denotes significant difference at $p < 0.05$).

<table>
<thead>
<tr>
<th></th>
<th>Successful REs</th>
<th>Successful REs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>NP*</td>
<td>7.35</td>
<td>3.958</td>
</tr>
<tr>
<td>VP</td>
<td>1.45</td>
<td>1.673</td>
</tr>
<tr>
<td>PRN</td>
<td>.02</td>
<td>.181</td>
</tr>
<tr>
<td>NNP*</td>
<td>.57</td>
<td>1.131</td>
</tr>
<tr>
<td>NN*</td>
<td>4.2</td>
<td>2.284</td>
</tr>
<tr>
<td>DT</td>
<td>2.59</td>
<td>1.791</td>
</tr>
<tr>
<td>JJ*</td>
<td>1.92</td>
<td>1.61</td>
</tr>
<tr>
<td>CC</td>
<td>.4</td>
<td>.645</td>
</tr>
<tr>
<td>PP</td>
<td>2.52</td>
<td>1.754</td>
</tr>
<tr>
<td>ADJP</td>
<td>.2</td>
<td>.536</td>
</tr>
<tr>
<td>ADVH</td>
<td>.27</td>
<td>.538</td>
</tr>
<tr>
<td>RB</td>
<td>.34</td>
<td>.639</td>
</tr>
<tr>
<td>VBN*</td>
<td>.22</td>
<td>.465</td>
</tr>
<tr>
<td>NNS</td>
<td>.61</td>
<td>.902</td>
</tr>
<tr>
<td>CD</td>
<td>.27</td>
<td>.552</td>
</tr>
</tbody>
</table>

Table 4: Models and their fit.

<table>
<thead>
<tr>
<th>Models</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic: NP+PP+ADVP+CD+length</td>
<td>.15</td>
</tr>
<tr>
<td>Semantic model: taxonomic + absolute</td>
<td>.338</td>
</tr>
<tr>
<td>Joint model: PP + taxonomic + absolute</td>
<td>.407</td>
</tr>
</tbody>
</table>

Predicting Image Success Rate ($\overline{NSR}_{image}$): We repeat a similar analysis for estimating how syntactic features relate to image success rate, i.e. how the image complexity, as estimated from the success rate of an image, influences how humans describe the target object, i.e. how human generated descriptions change with respect to the image complexity as estimated from the (normalised) success rate of an image. We find that humans use significantly more PPs and number of words ($p < 0.05$) when describing complex images.

In sum, syntactical features, which further describe and distinguish the target object (such as NPs, ADJ, and ADVPs and PPs) indicate successful REs. However, they cannot fully answer the question of “what makes a RE successful”, therefore we enrich our feature set using manually annotated semantic features.

4.2 Semantic Analysis of REG
We extract semantic features by annotating spatial frames of reference as described in (Gargett et al., 2010). We annotate a sample of the corpus (100 instances), which allows us to perform a direct comparison between the two corpora.

Comparison to GIVE-2 Corpus: We observe that in the REAL corpus, the taxonomic property, the relative property and the macro-level landmark
Table 5: Frequency of semantic frames in REAL vs. GIVE-2 (* denotes significant differences at \( p < 0.05, \chi^2 \) test).

\( \text{intrinsic property} \) of the object in question are used significantly more often than in the GIVE-2 corpus (Table 5).^{2}

In contrast, in GIVE-2 the \textit{absolute property} of the object, such as the colour, and references to \textit{distractors} are used significantly more often than in REAL. These results reflect the fact that scenes in REAL were more complex, and as such, relative properties to other objects and landmarks were used more often. In GIVE-2, target objects were mostly buttons, therefore, absolute descriptions (“the blue button”) or referring to an intrinsic distractor (“the red button next to the green”) are more frequent. In addition, real-world environments are dynamic. Humans choose to refer to immovable objects (\textit{macro-level landmarks}) more often than in closed-world environments. In GIVE-2, immovable objects are limited to walls, ceilings or floors, whereas in REAL there is a wide range of immovable objects /landmarks that a user can refer to, e.g. another building, rivers, parks, shops, etc. Landmark descriptions will play an important role in future navigation systems (Kandagath and Tu, 2015).

**Predicting RE Success Rate** \( (\widehat{S_{RE}}) \): Next, we analyse which spatial frames significantly contribute to task success, using \textit{multiple step-wise linear regression}. We find that \textit{taxonomic} and \textit{absolute} properties significantly \( (p < 0.05) \) contribute to the success of a referring expression (Table 4). Semantic features explain more of the variance observed in \( \widehat{S_{RE}} \), than syntactic features.

4.3 Joint Model of REG Success

Both syntactic and semantic features contribute to the success of a RE. Therefore, we construct a joint model for predicting \( \widehat{S_{RE}} \) using \textit{step-wise linear regression} over the joint feature space. We find that both syntactic and semantic features significantly \( (p < 0.05) \) contribute to the success of a RE, see Table 4. This model explains almost half of the variation observed in \( \widehat{S_{RE}} (R^2 = .407) \).

Clarke et al. (2013) reports an influence of visual salience on REG, therefore, in future, we will investigate the influence of visual features.

5 Discussion and Conclusions

From the results presented, the following conclusions can be drawn for real-world NLG systems. Firstly, semantic features have a bigger impact on the success rate of REs than syntactic features, i.e. content selection is more important than surface realisation for REG. Secondly, semantic features such as \textit{taxonomic} and \textit{absolute} properties can significantly contribute to RE success. Taxonomic properties refer to the type of target object, and in general depend on the local knowledge of the information giver. Similarly, the success of the RE will depend on the expertise of the information follower. As such, modelling the user’s level of knowledge (Janarthenam et al., 2011) and stylistic differences (Di Fabbrizio et al., 2008) is crucial. Absolute properties refer to object attributes, such as colour. Attribute selection for REG has attracted a considerable amount of attention, therefore it would be interesting to investigate how these automatic attribute selection algorithms perform in real-world, interactive environments. Finally, the more complex scenes seem to justify longer and more complex descriptions. As such, there is an underlying trade-off which needs to be optimised, e.g. following the generation framework described in (Rieser et al., 2014).

In future, we will compare existing REG algorithms on our dataset, in a similar experiment to Mitchell et al. (2013b). Then, we will extend existing algorithms to take into account other properties such as material (e.g. “wooden”), components of the referred object (e.g. “balconies”) etc. Finally, we will incorporate such an algorithm in interactive settings to investigate the influence of user dialogue behaviour and the influence of visual features, such as salience (Clarke et al., 2013), in order to improve the fit of our predictive model.
Acknowledgments

This research received funding from the EPSRC GUI project Generation for Uncertain Information (EP/L026775/1). The data has been collected through the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (SPACEBOOK project).

References


Srinivasan Janarthanam, Helen Hastie, Oliver Lemon, and Xingkun Liu. 2011. "the day after the day after tomorrow?" a machine learning approach to adaptive temporal expression generation: training and evaluation with real users. In 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue.


