The REAL Corpus: A Crowd-Sourced Corpus of Human Generated and Evaluated Spatial References to Real-World Urban Scenes

Phil Bartie[♢], **William Mackaness**[♠], **Dimitra Gkatzia**[⋆], **Verena Rieser**[⋆]
[♢]School of Natural Sciences, [♠]School of GeoSciences, [⋆]Department of Computer Science
University of Stirling, University of Edinburgh, Heriot-Watt University
phil.bartie@stir.ac.uk, william.mackaness@ed.ac.uk, {d.gkatzia,v.t.rieser}@hw.ac.uk

Abstract

We present a newly crowd-sourced data set of natural language references to objects anchored in complex urban scenes (In short: The REAL Corpus – Referring Expressions Anchored Language). The REAL corpus contains a collection of images of real-world urban scenes together with verbal descriptions of target objects generated by humans, paired with data on how successful other people were able to identify the same object based on these descriptions. In total, the corpus contains 32 images with on average 27 descriptions per image and 3 verifications for each description. In addition, the corpus is annotated with a variety of linguistically motivated features. The paper highlights issues posed by collecting data using crowd-sourcing with an unrestricted input format, as well as using real-world urban scenes. The corpus will be released via the ELRA repository as part of this submission.

Keywords: Image Descriptions, Spatial Referring Expressions, Urban Scenes, Vision and Language

1. Introduction

Generating successful 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 corpus of spatial references from real scenes rather than virtual.

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., 2013; 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.

2. 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) using a DSLR with a wide angle lens. The images were captured very early one summer morning 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 188 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). In order to encourage people to contribute more data, we added their email address to a prize draw for a £50 Amazon voucher, with an additional entry added for each expression they generated.

2.2. Collected Data

Overall, 873 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 2617 image tag verifications. Table 1 summarises the collected data.

# participants	188
# images/ stimuli	32
# descriptions	873
# verifications	2617

Table 1: Data in the REAL corpus

The type of data collected is notably different from previous corpus-based work on REG. Previous work has focused on highly controlled environments, such as virtual environments as used 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 infeasible. Consider, for instance, the building highlighted in Figure 2, for which the following descriptions were collected:

Figure 1: Original picture.

Figure 2: Target object in yellow box.

Figure 3: Identified objects by validators (green and red dots).

1. *A Greek style building with 5 pillars behind the first building that looks very similar. Left side of the image.*
2. *National art gallery of Scotland.*
3. *A classic building, with columns and a triangular pediment.*
4. *Building with columns behind the one in front, also in similar architecture style.*
5. *Building with parked cars in front.*
6. *The building of interest is the one on the far left of the image, the second building after the front and big one. Seems like an old building.*
7. *Building behind the main building. Both have columns on the side facing the camera.*
8. *The second building with columns, the farthest away one.*
9. *The faraway set of pillars, on the other side of the national gallery.*
10. *The back end of the National Galleries building that is furthest away.*
11. *The columned building in front of the round green copper coloured rooftop, which is left of the steeple on the skyline.*

It is evident that the REAL users refer to a variety of object qualities. We observe that some participants refer to the architecture of the building explicitly or implicitly (*similar architecture style, Greek style building, classic building, triangular pediment*), some refer to movable objects (*parked cars in front*) and some make use of the location (*second building, in front of*).

2.3. Corpus Annotation

Information on participants: The web interface first asked participants to enter information on their age group and gender. The corpus contains data from 90 male and 98 female participants. The age groups are distributed as shown in Table 2. Over half of the participants are aged between 21 and 30 years old, but all ages from 16 to 71 and older are represented.

Age group	Number of participants	Percentage
16 - 20	25	13.3%
21 - 30	105	55.85%
31 - 40	33	17.55%
41 - 50	15	8%
51 - 60	4	2.13%
61 - 70	4	2.13%
71 or older	2	1.1%

Table 2: Distribution of age groups

Syntactic Features: We use the Stanford CoreNLP tool (Manning et al., 2014) to syntactically annotate the human generated REs. In previous work (Gkatzia et al., 2015) we found that the following syntactic categories predict successful REs: NP (Noun phrases), NNP (Proper noun, singular), NN (Noun, singular or mass), JJ (Adjective) and VBN (Verb, past participle). For example, the following descriptions uses NNPs and NNs to distinguish the reference object:

> *The **large American-style wooden building** with **balcony painted cream and red/brown**. **Ground floor** is a **cafe with tables and parasols** outside.*

Semantic Features: We also manually annotated a sample of 100 corpus instances with semantic features using spatial frames of reference as described in (Gargett et al., 2010), see Table 3.

Human Identification Success Rates: In order to verify the human generated RE, the respondent clicked on the image where they believed the target to be based on the description. They were also able to respond with "ambiguous" if they considered there to be more than one matching object in the scene, or "not found" if they were unable to find any suitable object based on the description given. All cases where the respondent clicked on the image were manually checked to determine if the 'correct' (green) or 'incorrect' (red) target had been identified Fig. 3. Overall, 77.5% of human descriptions provided were successfully identified. Also see Table 4.

RE Success Rates: In previous work (Gkatzia et al., 2015) we have used the REAL corpus to automatically predict the success of REs[1]. In particular, the corpus is annotated with the following measures of success:

[1]The Java code for the normalisation of the success rate can be found at `https://github.com/dimi123/EMNLP-2015`

Type	Description	Example
Taxonomic property	Reference to the type of object that is aimed to be described.	*a coffeeshop*
Absolute property	Reference to a property/ attribute, e.g. colour, of the object that can be determined without comparing it to other objects.	*the white building*
Relative property	Reference to a property of the object in relation to other similar objects.	*a tall building with large columns in the front*
Viewer-centred	Reference to the object's location relative to the viewer's location.	*The corner nearest to us on the right side of the road straight ahead, with a turret on top.*
Micro-level landmark intrinsic	Reference to the object's location in relation to a different movable object.	*there is a silver car parked in front*
Distractor intrinsic	Reference to the object's location in relation to another similar object (i.e. distractor).	*There is a building with similar apartments with red external stairs. The middle apartment that has a blue door in the first floor.*
Macro-level landmark intrinsic	reference to the object's location in relation to an immovable object.	*next to the river*
Deduction by elimination	reference to the object by specifying which objects are not meant and letting the viewer deduce the intended one.	*Look at Poundsavers. The big tall building not the red one to the left of it.*

Table 3: Manually annotated spatial frames, following (Gargett et al., 2010).

verification	total	success rate
ambiguous	249	9.5%
not found	84	3.2%
correct	2029	**77.5%**
incorrect	255	9.7%

Table 4: Success rates of human identification task

- **RE Success Rate**: Frequency of successful identification of the target object in verification phase per RE, i.e. average of 3 verifications per RE. This measure is used to estimate the quality of the human generated RE.
- **Image Success Rate**: Frequency of successful identification of the target object in verification phase per image, i.e. average of 27 verifications per image. This measure is used as an approximation for image complexity.

3. Release Format

The image stimuli, images with tag points (coloured red/green for correct/incorrect), referring expression data and corpus annotations as described in Section 2.3. will be released as part of this submission. Due to privacy restrictions the IP address and email address of participants will be withheld.

4. Conclusions and Future Work

We presented a dataset which consists of aligned images of real-world spatial scenes with accompanied referring expressions of specific objects in the images. Therefore, the dataset will be useful for research in the fields of referring expression generation, as well as language and vision. In addition, the dataset can be further used for paraphrasing due to several different ways to refer to objects. In the future, we aim to use the dataset for the development of algorithms for referring expressions generation.

Acknowledgments

This research received funding from the EPSRC projects GUI "Generation for Uncertain Information" (EP/L026775/1) and DILiGENt "Domain-Independent Language Generation"" (EP/M005429/1). The data has been collected through the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 216594 (SPACEBOOK project).

Fitzgerald, N., Artzi, Y., and Zettlemoyer, L. (2013). Learning distributions over logical forms for referring expression generation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gargett, A., Garoufi, K., Koller, A., and Striegnitz, K. (2010). The GIVE-2 Corpus of Giving Instructions in Virtual Environments. In *7th International Conference on Language Resources and Evaluation (LREC)*.

Gkatzia, D., Rieser, V., Bartie, P., and Mackaness, W. (2015). From the Virtual to the Real World: Referring to Objects in Real-World Spatial Scenes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1936–1942, Lisbon, Portugal, September. Association for Computational Linguistics.

Kazemzadeh, S., Ordonez, V., Matten, M., and Berg, T. (2014). ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Manning, C., Surdeanu, M., Finkel, J., Bethard, J., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *52nd Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*.

Mitchell, M., van Deemter, K., and Reiter, E. (2010). Natural reference to objects in a visual domain. In *6th International Natural Language Generation Conference (INLG)*.

Mitchell, M., Reiter, E., and van Deemter, K. (2013). Typicality and object reference. In *Cognitive Science (CogSci)*.

van Deemter, K., van der Sluis, I., and Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. In *4th International Natutral Language Generation Conference*.

Viethen, J. and Dale, R. (2008). The use of spatial relations in referring expression generation. In *5th International Natural Language Generation Conference (INLG)*.