Zoom: a corpus of natural language descriptions of map locations

Romina Altamirano FaMAF Nat.Univ. of Córdoba Córdoba, Argentina ialtamir@famaf.unc.edu.ar Thiago C. Ferreira EACH-USP Univ. of São Paulo São Paulo, Brazil thiago.castro.ferreira@usp.br

Abstract

This paper describes an experiment to elicit referring expressions from human subjects for research in natural language generation and related fields, and preliminary results of a computational model for the generation of these expressions. Unlike existing resources of this kind, the resulting data set - the Zoom corpus of natural language descriptions of map locations - takes into account a domain that is significantly closer to real-world applications than what has been considered in previous work, and addresses more complex situations of reference, including contexts with different levels of detail, and instances of singular and plural reference produced by speakers of Spanish and Portuguese.

1 Introduction

Referring Expression Generation (REG) is the computational task of producing adequate natural language descriptions (e.g., pronouns, definite descriptions, proper names, etc.) of domain entities. In particular, the issue of how to determine the semantic contents of definite descriptions (e.g., 'the Indian restaurant on 5th street', 'the restaurant we went to last night', etc.) has received significant attention in the field, and it is also the focus of the present work.

The input to a REG algorithm is a context set C containing an intended referent r and a number of distractor objects. All objects are represented as attribute-value pairs representing either *atomic* (*type-restaurant*) or *relational* (*on-5thstreet*) properties (Krahmer and Theune, 2002; Krahmer et al., 2003; Kelleher and Kruijff, 2006; Viethen et al., 2013). The expected output is a *uniquely identifying* list L of properties known to be true of r so that L distinguishes r from all distractors in C (Dale and Reiter, 1995).

Ivandré Paraboni EACH-USP Univ. of São Paulo São Paulo, Brazil ivandre@usp.br Luciana Benotti FaMAF Nat.Univ. of Córdoba Córdoba, Argentina benotti@famaf.unc.edu.ar

Properties are selected for inclusion in L according to multiple - and often conflicting - criteria, including discriminatory power (i.e., the ability to rule out distractors) as in (Dale, 2002; Gardent, 2002), domain preferences (Pechmann, 1989; Gatt et al., 2013) and many others. A description that conveys more information than what is strictly required for disambiguation is said to be *overspecified* (Arts et al., 2011; Koolen et al., 2011; van Gompel et al., 2012; Engelhardt and Ferreira, 2006; Engelhardt et al., 2011). For a review of the research challenges in REG, see (Krahmer and van Deemter, 2012).

Existing approaches to REG largely consist of algorithmic solutions, many of which have been influenced by, or adapted from, the Dale & Reiter Incremental algorithm in (Dale and Reiter, 1995). The use of machine learning (ML) techniques, by contrast, seems to be less frequent than in other NLG tasks, although a number of exceptions do exist (e.g., (Jordan and Walker, 2005; Viethen and Dale, 2010; Viethen, 2011; Garoufi and Koller, 2013; Ferreira and Paraboni, 2014)).

A possible explanation for the small interest in ML for REG may be the relatively low availability of data. While research in many fields may benefit from the wide availability of text corpora (e.g., obtainable from the web), research in REG usually requires highly specialised data - hereby called REG corpora - conveying not only referring expressions produced by human speakers, but also a fully-annotated representation of the context (i.e., all objects and their semantic properties) within which the expressions have been produced.

REG corpora such as TUNA (Gatt et al., 2007) and GRE3D3 (Dale and Viethen, 2009) are useful both to gain general insights on human language production, and to benefit from data-intensive computational techniques such as ML. However, being usually the final product of controlled experiments involving human subjects, REG corpora tend to address highly specific research questions. For instance, GRE3D3 is largely devoted to the investigation of relational referring expressions (Kelleher and Kruijff, 2006) in simple visual scenes involving geometric shapes, as in 'the large ball next to the red cube'. As a result, and despite the usefulness of these resources to a large body of work in REG, further research questions will usually require the collection of new data.

In this paper we introduce the Zoom corpus of referring expressions. Zoom addresses a domain that is considerably closer to real-world applications (namely, city maps in different degrees of detail represented by zoom levels) than what has been considered in previous work, involving both singular and plural reference, and making extensive use of relational properties. Moreover, Zoom descriptions were produced by both Spanish and Portuguese speakers, which will allow (to the best of our knowledge, for the first time) a comprehensive study of the REG surface realisation subtask in these languages, and enable research on the issues of human variation in REG (Fabbrizio et al., 2008; Altamirano et al., 2012; Gatt et al., 2011).

2 Related work

TUNA (Gatt et al., 2007) was the first prominent REG corpus to be made publicly available for research purposes. The corpus was developed in a series of controlled experiments, containing 2280 atomic descriptions produced by 60 speakers of English in two domains (1200 descriptions of furniture items and 1080 descriptions of people's photographs). TUNA has been used in a series of REG shared tasks (Gatt et al., 2009).

GRE3D3 and its extension GRE3D7 (Dale and Viethen, 2009; Viethen and Dale, 2011) were developed in a series of web-based experiments primarily focussed on the study of relational descriptions. GRE3D3 contains 630 descriptions produced by 63 speakers, and GRE3D7 contains 4480 descriptions produced by 287 speakers. In both cases, the language of the experiment was English. The domain consists of simple visual scenes conveying boxes and spheres.

Stars (Teixeira et al., 2014) and its extension Stars2 were collected for the study of referential overspecification. Stars contains 704 descriptions produced by 64 speakers in a web-based experiment. Stars2 was produced in dialogue situations involving subject pairs, and it contains 884 descriptions produced by 56 speakers. Both domains make use of simple visual scenes containing up to four object types (e.g., stars, boxes, cones and spheres) and include atomic and relational descriptions alike. The language of both experiments was Brazilian Portuguese.

3 Experiment

We designed a web-based experiment to collect natural language descriptions of map locations in both Spanish and Portuguese. The collected data set comprises a corpus of referring expressions for research in REG and related fields. The situations of reference under consideration make use of map scenes in two degrees of detail (represented by low and high zoom levels), and address instances of singular and plural reference. A fragment of the experiment interface is shown in Fig. 1.

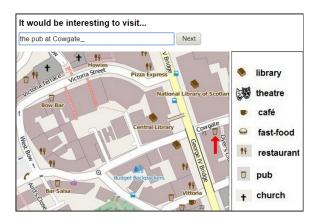


Figure 1: Experiment interface

3.1 Subjects

Volunteers were recruited upon invitation sent by email. The Portuguese data had 93 participants, being 66 (71.0%) male and 27 (29.0%) female. The Spanish data had 80 participants, being 59 male (69.4%) and 26 female (30.6%).

3.2 Procedure

Subjects received a web link to the on-line experiment interface (cf. Fig. 1) with self-contained instructions. Age and gender details were collected for statistical purposes. The experiment consisted of a series of map images presented in random order, one by one. Each map scene showed a particular location (e.g., a restaurant, pub, theatre etc.) pointed by an arrow. For each scene, subjects were required to imagine that they were giving travel advice to a friend, and to complete the sentence 'It would be interesting to visit...' with a description of the location pointed by the arrow. After pressing a 'Next' button, another stimulus was selected, until the end of the experiment. The first two images were fillers solely intended to make subjects familiar with the experiment setting, and the corresponding responses were not recorded. Incomplete trials, and ill-formed descriptions, were also discarded.

3.3 Materials

The experiment made use of the purpose-built interface illustrated in Fig. 1, and a set of map images obtained from OpenStreetMap¹, which consisted of selected portions of maps of Madrid and Lisbon to be presented to Spanish and Portuguese speakers, respectively. For each city, 10 map locations were used. Each location was shown in low and high zoom levels, making 20 images in total. In both cases, the intended target was kept the same, but the more detailed version would display a larger number of distractors and additional details in general. In addition to that, certain street and landmark names might not be depicted at different zoom levels. Half images showed a single arrow pointing to one map location (i.e., requiring a single description as 'the restaurant on Baker street'), whereas the other half showed two arrows pointed to two different locations (and hence requiring a reference to a set, as in 'the two restaurants near the museum').

3.4 Data collection

Upon manual verification, 602 ill-formed Portuguese descriptions and 366 Spanish descriptions were discarded. Thus, the Portuguese subcorpus consists of 1358 descriptions, and the Spanish subcorpus consists of 1234 descriptions. In the Portuguese subcorpus, 78.6% of the descriptions include relational properties. In addition to that, 36.4% were minimally distinguishing, 44.3% were overspecified, and 19.3% were underspecified. In the Spanish subcorpus, 70% of the descriptions include relational properties, 35% were minimally distinguishing, 40% were overspecified, and 25% were underspecified. Underspecified descriptions are not common in existing REG corpora (i.e., certainly not in this proportion), which may reflect the complexity of the domain and/or limitations of the web-based setting.

3.5 Annotation

Each referring expression was modelled as conveying a description of the main target object and, optionally, up to four descriptions of related landmarks. The annotation scheme consisted of three target attributes, four landmark attributes for each of the four possible landmark objects, and seven relational properties. This makes 26 possible attributes for each referring expression. In the case of plural descriptions (i.e., those involving two target objects), this attribute set is doubled.

Every object was annotated with the atomic attributes *type*, *name* and *others* and, in the case of landmark objects, also with their *id*. In addition to that, seven relational properties were considered: *in/on/at*², *next-to*, *right-of*, *left-of*, *in-frontof*, *behind-of*, and the multivalue relation *between* intended to represent 'corner' relations.

Possible values for the *type* and *name* attributes are predefined by each referential context. The *others* attribute may be assigned any string value, and it is intended to represent any non-standard piece of information conveyed by the expression. For the spatial relations, possible values are the object identifiers available from each scene.

The collected descriptions were fully annotated by two independent annotators. After completion, a third annotator assumed the role of judge and provided the final annotation. Since the annotation scheme was fairly straightforward (i.e., largely because all non-standard responses were simply assigned to the *others* attribute), agreement between judges as measured by Kappa (Cohen, 1960) was 84% at the attribute level. Both referential contexts and referring expressions were represented in XML format using a relational version of the format adopted in TUNA (Gatt et al., 2007).

3.6 Comparison with previous work

Table 1 presents a comparison between the collected data and existing REG corpora³: the number of referring expressions (REs), the number of subjects in each experiment, the number of possible atomic attributes (Attrib.) and possible landmarks (LMs) in a description, the average description size (in number of annotated properties), and the proportion of property usage, which is taken to

¹openstreetmap.org

²The three prepositions were aggregated as a single attribute because they have approximately the same meaning in the languages under consideration

³The information on TUNA and Zoom descriptions is based on the singular portion of each corpus only

be the proportion of properties that appear in the description over the total number of possible attributes and landmarks. From a REG perspective, larger description sizes and lower usage rates may suggest more complex situations of reference.

Table 1: Comparison with existing REG corpora

Corpus	REs	Subj.	Attrib.	LMs	Avg.size	Usage
TUNA-F	1200	60	4	0	3.1	0.8
TUNA-P	1080	60	10	0	3.1	0.3
GRE3D3	630	63	9	1	3.4	0.3
GRE3D7	4480	287	6	1	3.0	0.4
Stars	704	64	8	2	4.4	0.4
Stars2	884	56	9	2	3.3	0.3
Zoom-Pt	1358	93	19	4	6.7	0.3
Zoom-Sp	1234	80	19	4	7.2	0.3

4 **REG evaluation**

In what follows we illustrate the use of the Zoom corpus as training and test data for a simple machine learning approach to REG adapted from (Ferreira and Paraboni, 2014). The goal of this evaluation is to provide reference results for future comparison with purpose-built REG algorithms, and not to present a complete REG solution for the Zoom domain or others.

The present model consists of 12 binary classifiers representing whether individual referential attributes should be selected for inclusion in an output description. The classifiers correspond to atomic attributes of the target and first landmark object (*type*, *name* and *others*), and relations. Referential attributes of other landmark objects were not modelled due to data sparsity and also to reduce computational costs. For similar reasons, the multivalue *between* relation is also presently disregarded, and 'corner' relations involving two landmarks (e.g., two streets) will be modelled as two independent classification tasks.

Only two learning features are considered by each classifier: *landmarkCount*, which represents the number of landmark objects near the main target, and *distractorCount*, which represents the number of objects of the same type as the target within the relevant context in the map. For other possible features applicable to this task, see, for instance, (dos Santos Silva and Paraboni, 2015).

From the outcome of the 12 binary classifiers, a description is built by considering atomic target attributes in the first place. All attributes that correspond to a positive prediction are selected for inclusion in the output description. Next, relations are considered. If no relation is predicted, the algorithm terminates by returning an atomic description of the main target object. If the description includes a relation, the corresponding landmark object is selected, and the algorithm is called recursively to describe it as well. Since every attribute that corresponds to a positive prediction is always selected, the algorithm does not regard uniqueness as a stop condition. As a result, the output description may convey a certain amount of overspecification.

For evaluation purposes, we used the subset of singular descriptions from the Portuguese portion of the corpus, comprising 821 descriptions. Evaluation was carried out by comparing the corpus description with the system output to measure overall accuracy (i.e., the number of exact matches between the two descriptions), Dice (Dice, 1945) and MASI (Passonneau, 2006) coefficients.

Following (Ferreira and Paraboni, 2014), we built a REG model using support vector machines with radial basis function kernel. The classifiers were trained and tested using 6-fold cross validation. Optimal parameters were selected using grid search as follows: for each step in the main k-fold validation, one fold was reserved for testing, and the remaining k - 1 folds were subject to a secondary cross-validation procedure in which different parameter combinations were attempted. The C parameter was assigned the values 1, 10, 100 and 1000, and γ was assigned 1, 0.1, 0.001 and 0.0001. The best-performing parameter set was selected to build a classifier trained from the k-1fold, and tested on the test data. This was repeated for every iteration of the main cross-validation procedure.

Table 2 summarises the results obtained by the REG algorithm built from SVM classifiers, those obtained by a baseline system representing a relational extension of the Dale & Reiter Incremental Algorithm, and by a Random selection strategy.

Table 2: REG results						
Algorithm	Acc.	Dice	MASI			
SVM	0.15	0.51	0.28			
Incremental	0.04	0.53	0.21			
Random selection	0.03	0.45	0.15			

We compare accuracy scores obtained by every algorithm pair using the chi-square test, and we compare *Dice* scores using *Wilcoxon's* signedrank test. In terms of overall accuracy, the SVM approach outperforms both alternatives. The difference from the second best-performing algorithm (i.e., the Incremental approach) is significant ($\chi^2 = 79.87$, df=1, p<0.0001). Only in terms of Dice scores a small effect in the opposite direction is observed (T=137570.5, p= 0.01413).

We also assessed the performance of the individual classifiers. Table 3 shows these results as measured by precision (P), recall (R), F1-measure (F1) and area under the ROC curve (AUC).

Table 3: Classifier results

Classifier	Р	R	F_1	AUC
tg_type	0.95	1.00	0.98	0.25
tg₋name	0.09	0.05	0.07	0.41
tg_other	0.00	0.00	0.00	0.05
lm_type	0.93	1.00	0.96	0.44
lm_name	0.97	1.00	0.98	0.35
lm_other	0.00	0.00	0.00	0.43
next-to	0.50	0.24	0.32	0.63
right-of	0.00	0.00	0.00	0.28
left-of	0.00	0.00	0.00	0.27
in-front-of	0.00	0.00	0.00	0.42
behind-of	0.00	0.00	0.00	0.17
in/on/at	0.60	0.60	0.60	0.61

From these results we notice that highly frequent attributes (e.g., target type and landmark name) were classified with high accuracy, whereas others (e.g., multivalue attributes and relations) were not.

5 Discussion

This paper has introduced the Zoom corpus of natural language descriptions of map locations, a resource intended to support future research in REG and related fields. Preliminary results of a SVMbased approach to REG - which were solely presented for the future assessment of REG algorithms based on Zoom data - hint at the actual complexity of the REG task in this domain in a number of ways. First, we notice that a similar approach in (Ferreira and Paraboni, 2014) on GRE3D3 and GRE3D7 data has obtained considerably higher mean accuracy. This is partially explained by the increased complexity of the Zoom domain, but also by the currently simple annotation scheme.

Second, we notice that Zoom descriptions are prone to convey relations between a single target and multiple landmark objects, as in 'the restaurant between the 5th and 6th streets'. Although common in language use, the use of multiple relational properties in this way has been little investigated in the REG field.

Finally, we notice that the Zoom domain contains two descriptions for every target object, which are based on different - but related - models corresponding to the same map location seen at different zoom levels. Interestingly, the referring expression in a 1X situation may or may not be the same as in a 2X situation. Consider a map with higher zoom level (2X) as illustrated in the previous Fig. 2, and the same map location as seen with lower zoom level in the previous Fig. 1.



Figure 2: Map with a more detailed zoom level

The underlying models for these two maps are certainly different, but not unrelated. The map with 2X zoom contains fewer objects but may include more properties due to the added level of detail. The referring expression for the target in the 1X map may or may not be the same as in the 2X map. For instance, the referring expression "the pub at Cowgate" is underspecified on the 1X map, but it is minimally distinguishing on the 2X map.

Differences of this kind are common in interactive applications (e.g., in which the context of reference may change in structure or in the number of objects and referable properties), and the challenge for REG algorithms would be to produced an appropriate description for the modified context without starting from scratch. REG algorithms based on local context partitioning (Areces et al., 2008) may have an advantage in this respect, but further investigation is still required.

Acknowledgements

This work has been supported by FAPESP and CAPES. The authors are also grateful to the annotators Adriano S. R. da Silva, Jefferson S. F. da Silva and Alan K. Yamasaki.

References

- Romina Altamirano, Carlos Areces, and Luciana Benotti. 2012. Probabilistic refinement algorithms for the generation of referring expressions. In *COL-ING (Posters)*, pages 53–62.
- Carlos Areces, Alexander Koller, and Kristina Striegnitz. 2008. Referring expressions as formulas of description logic. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 42–49, Stroudsburg, PA, USA. Association for Computational Linguistics.
- A. Arts, A. Maes, L. G. M. Noordman, and C. Jansen. 2011. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.
- J. Cohen. 1960. A coeficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.
- Robert Dale and Jette Viethen. 2009. Referring expression generation through attribute-based heuristics. In *Proceedings of ENLG-2009*, pages 58–65.
- Robert Dale. 2002. Cooking up referring expressions. In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, pages 68–75.
- L. R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.
- Diego dos Santos Silva and Ivandré Paraboni. 2015. Generating spatial referring expressions in interactive 3D worlds. *Spatial Cognition and Computation*.
- P. E. Engelhardt and K. Baileyand F. Ferreira. 2006. Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573.
- P. E. Engelhardt, S. B. Demiral, and Fernanda Ferreira. 2011. Over-specified referring expressions impair comprehension: An ERP study. *Brain and Cognition*, 77(2):304–314.
- Giuseppe Di Fabbrizio, Amanda J. Stent, and Srinivas Bangalore. 2008. Trainable speaker-based referring expression generation. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, CoNLL '08, pages 151–158, Stroudsburg, PA, USA.
- Thiago Castro Ferreira and Ivandré Paraboni. 2014. Classification-based referring expression generation. *Lecture Notes in Computer Science*, 8403:481– 491.

- C. Gardent. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 96–103.
- Konstantina Garoufi and Alexander Koller. 2013. Generation of effective referring expressions in situated context. *Language and Cognitive Processes*, 29(8):986–1001.
- Albert Gatt, Ilka van der Sluis, and Kees van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of ENLG-07*.
- Albert Gatt, Anja Belz, and Eric Kow. 2009. The TUNA challenge 2009: Overview and evaluation results. In *Proceedings of the 12nd European Workshop on Natural Language Generation*, pages 174–182.
- Albert Gatt, R. van Gompel, E. Krahmer, and K. van Deemter. 2011. Non-deterministic attribute selection in reference production. In Workshop on the Production of Referring Expressions (PRE-CogSci 2011), pages 1–7.
- Albert Gatt, E. Krahmer, R. van Gompel, and K. van Deemter. 2013. Production of referring expressions: Preference trumps discrimination. In 35th Meeting of the Cognitive Science Society, pages 483–488.
- Pamela W. Jordan and Marilyn A. Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. J. Artif. Int. Res., 24(1):157–194.
- J. D. Kelleher and G. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pages 1041–1048.
- 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 Mariet Theune. 2002. Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, pages 223–264. CSLI Publications, Stanford, CA.
- Emiel Krahmer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.
- Emiel Krahmer, Sebastiaan van Erk, and Andre Verleg. 2003. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72.

- Rebecca Passonneau. 2006. Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC).*
- T. Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27(1):98–110.
- Caio V. M. Teixeira, Ivandré Paraboni, Adriano S. R. da Silva, and Alan K. Yamasaki. 2014. Generating relational descriptions involving mutual disambiguation. *Lecture Notes in Computer Science*, 8403:492–502.
- R. van Gompel, Albert Gatt, E. Krahmer, and K. van Deemter. 2012. PRO: A computational model of referential overspecification. In *Proceedings of AMLAP-2012*.
- Jette Viethen and Robert Dale. 2010. Speakerdependent variation in content selection for referring expression generation. In *Proceedings of the Australasian Language Technology Association Workshop 2010*, pages 81–89, Melbourne, Australia.
- Jette Viethen and Robert Dale. 2011. GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of UCNLG+Eval-2011*, pages 12–22.
- Jette Viethen, Margaret Mitchell, and Emiel Krahmer. 2013. Graphs and spatial relations in the generation of referring expressions. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 72–81, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Jette Viethen. 2011. The Generation of Natural Descriptions: Corpus-Based Investigations of Referring Expressions in Visual Domains. Ph.D. thesis, Macquarie University, Sydney, Australia.