Referring Expression Generation: Taking Speakers' Preferences into Account

Thiago Castro Ferreira and Ivandré Paraboni

School of Arts, Sciences and Humanities, University of São Paulo (EACH / USP)
Av. Arlindo Bettio, 1000 - São Paulo, Brazil
{thiago.castro.ferreira,ivandre}@usp.br

Abstract. We describe a classification-based approach to referring expression generation (REG) making use of standard context-related features, and an extension that adds speaker-related features. Results show that taking speakers' preferences into account outperforms the standard REG model in four test corpora of definite descriptions.

Keywords: Natural Language Generation, Referring Expressions

1 Introduction

In Natural Language Generation (NLG), Referring Expression Generation (REG) is the computational task of uniquely describing a given target object *r* within its context *C*.[1] For instance, given a visual context as in Figure 1, we may compute descriptions of a particular entity *r* as 'the cone next to a box', 'the cone closest to the centre' etc., among other possibilities.

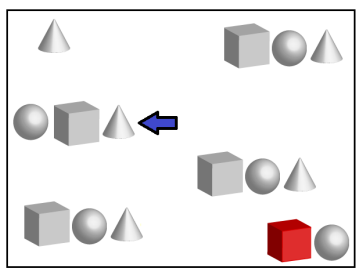

Fig. 1. A simple visual context, with a target object pointed by the arrow.

The input to the REG task is a set of context objects *C* (including the intended referent *r*) and their possible semantic properties represented as attribute-value pairs, as in (*type-cone*) or (*colour-grey*). The goal is to produce a list *L* of properties of *r* so as to distinguish *r* from all other objects (or distractors) in *C* [2]. Moreover, as in many

[1] We presently focus on content selection. For surface realisation issues, see, e.g., [11].

P. Sojka et al. (Eds.): TSD 2014, LNAI 8655, pp. 539–546, 2014.
This is preprint prepared by Proceedings editor for Springer International Publishing Switzerland.

NLG tasks, it is generally assumed that *L* should reflect the human choice of contents as closely as possible.

REG implementations range from purely algorithmic proposals (e.g., [2,10]), to the more recent use of machine learning techniques [5]. Regardless of implementation, however, the output description tends to be the same (i.e., fixed) for each input context, and this is so despite evidence that speakers may differ from each other in their choices of reference strategy. In other words, many existing REG algorithms do not pay regard to the fact that different people may refer to the same object in different ways.

Differences across speakers are observed in many aspects of language production. For instance, speakers may vary the choices of referential attributes, or the degree of reference specification. Regarding attribute choice, the work in [13], for instance, describes an experiment in which 15% of speakers always made use of relational descriptions (e.g., 'the cone *next to* a box'), whereas 31% never used them (e.g., favouring atomic properties as in 'the *middle* one ').

As for reference specification, the experiment in [12], for instance, shows that 12% of the speakers always chose a minimal description (e.g., 'the cone next to the box'), whereas 16% always overspecified (e.g., 'the cone next to the *grey* box', in a context in which colour was redundant).

The reasons why reference strategies may vary across speakers are beyond the scope of this work. In what follows, we simplify and assume that these differences represent linguistic preferences at some level, and we focus instead on the use of machine learning techniques that take advantage of speaker-related information, an issue that few approaches to REG have taken into account to date.

2 Background

The NLG literature presents a wide range of approaches to REG that are mainly based on the Incremental algorithm [2], some of which are described in [9]. Studies that take speaker-related information into account are summarized below.

The work in [6] describes an experiment using the Incremental algorithm and one of its extensions applied to the generation of descriptions in two dialogue corpora. Attributes are selected in order of recency per speaker. The speaker-dependent version is shown to outperform the standard Incremental algorithm with a fixed preference list for attribute selection.

In [1,4], attributes are selected based on their frequency per speaker. In both studies, the use of speaker-related information is once again shown to outperform the Incremental algorithm with a fixed preference list. The work in [4] presents also a speaker-dependent REG strategy that computes all possible descriptions of a given target, and selects either the most frequent or most recent form for each speaker. In both cases, using speaker-related information outperforms a standard algorithm that selects the most frequent form in the training data.

The more recent availability of larger sets of referring expressions and their accompanying contexts (or REG corpora) has enabled a number of corpus-based approaches to REG as well. Among these, of special interest to our present work are *GRE3D3* and

GRE3D7 [13,15], *Stars* [12] and *Stars2* [7] corpora. Each of these corpora were produced by a relatively large number of human participants, making them more likely to show variation across speakers. *GRE3D3* [13] and *GRE3D7* [15] concern descriptions of objects in simple 3D scenes, which may be either atomic (e.g., 'the red box') or involve a spatial relation (e.g., 'the red box on top of the cube'). *GRE3D3* contains 630 descriptions produced by 63 speakers, and *GRE3D7* contains 4480 descriptions produced by 287 speakers.

Stars [12] and *Stars2* [7] concern descriptions in 2D scenes involving up to three objects each (namely, a target, first and second landmark objects, as in 'the cube next to the cone below the red sphere'). Previous Figure 1 illustrates a scene from the *Stars2* corpus. *Stars* contains 704 descriptions produced by 64 speakers, and *Stars2* [7] contains 1216 descriptions produced by 76 speakers.

The availability of corpus knowledge allows the use of machine learning techniques in REG. For instance, the work in [14] applies decision-tree induction to predict reference strategies. In one of their experiments, the REG model includes information that identifies each speaker, which generally outperforms alternatives that did not take speaker information into account.

The work in [5] applies support vector machine (SVM) classifiers to the REG task. Results show that the SVM approach outperformed a standard implementation of the Incremental algorithm on both *GRE3D3* and *GRE3D7* corpora. Our present work follows the SVM approach in [5], which is presently expanded in a number of ways: by considering additional context features, additional corpora for training and evaluation purposes, and a set of speaker-related features.

3 Current Work

As in [5], we make use of a classification-based approach to REG built on support vector machines with radial basis function kernel. Two kinds of classifier are considered: binary classifiers for each individual atomic attribute prediction (e.g., colour, size etc.), and multi-class classifiers for relational attribute prediction (e.g., the choice between left(x), above(x) etc. where x is a landmark object.)

The classifiers for each corpus may be related to a target, first or second landmark object mentioned in each description. For instance, 'the box next to the sphere below the cone' conveys a reference to a target (box), a first (sphere) and second (cone) landmarks. Given our goal to evaluate REG algorithms in four corpora of referring expressions (namely, *GRE3D3*, *GRE3D7*, *Stars* and *Stars2*, cf. previous section), the number of classifiers is slightly different for each domain. For instance, attributes representing vertical and horizontal screen position are only relevant to the 2D scenes in *Stars* and *Stars2* data, but not to *GRE3D3* and *GRE3D7* 3D scenes. Furthermore, *Stars* and *Stars2* descriptions may refer to up to two landmark objects, whereas for *GRE3D3* and *GRE3D7* descriptions there is a one landmark maximum.

GRE3D3 and *GRE3D7* descriptions required 8 binary classifier for individual attribute prediction, and one multi-class classifier for the relation attribute prediction. Possible values for the relation attribute in this case are *no relation*, *right-of*, *left-of*, *next-to*, *on-top-of* and, in the case of *GRE3D3*, also *in-front-of*.

Stars descriptions required 12 binary classifier for individual attribute prediction, and *Stars2* descriptions required 15. In both cases, two multi-class classifiers were trained: one for the relation attribute between the target and the first landmark, and another for the relation between the first and second landmarks. Possible values for the relation attributes are *no relation*, *right-of*, *left-of*, *next-to*, *below*, *above* and, in the case of *Stars2*, also *in-front-of*.

From the set of binary and multi-class classifiers for each individual attribute, a description is built as follows. First, atomic target attributes are considered. For each positive class prediction, the corresponding attribute will be included in the target description. Next, relational attributes are considered. If no relational attribute is predicted, the algorithm terminates by returning an atomic description L of the target. On the other hand, if L involves a relation then the related landmark object is included in the output description and the algorithm is called once again to describe it recursively.

The algorithm does not explicitly test for uniqueness of the output description under generation, that is, every attribute that corresponds to a positive class is always included in L, which in many cases will become overspecified. This, as we shall see in the results described in Section 4, turns out to produce descriptions that resemble those produced by human speakers in the test data.

As most existing REG algorithms, the current model - which is solely based on context-related features - will always produce the same (i.e., fixed) output description for a given target input. This model, hereby called $-SP$, will be used as a standard, speaker-independent REG approach. The context features used by this model are summarized in Table 1. In the *Stars* corpus, the features related to the size of objects (TG_Size, LM_Size, Num_TG_Size, Num_LM_Size, TG_LM_Same_Size) are not used.

Table 1. Context-related features, taken from [14]

Feature	Description
TG_Size	target size
LM_Size	landmark size
Relation_Type	type of relation between target and landmark
Num_TG_Size	number of objects of same size as the target
Num_LM_Size	number of objects of same size as landmark
TG_LM_Same_Size	target and landmark share size
Num_TG_Col	number of objects of same colour as target
Num_LM_Col	number of objects of same colour as landmark
TG_LM_Same_Col	target and landmark share colour
Num_TG_Type	number of objects of same type as target
Num_LM_Type	number of objects of same type as landmark
TG_LM_Same_Type	target and landmark share type

In addition to that, we will also consider an expanded version of the model containing a number of speaker-related features. This speaker-dependent model will be called $+SP$, and conveys, besides the information in $-SP$, personal and attribute usage information as seen in the training data. These features are summarized in Table 2.

Table 2. Speaker-related Features

Feature	Description
Speaker_ID	speaker's unique identifier
Speaker_Gender	speaker's gender
Speaker_AgeGroup	speaker's age group
Speaker_Frequency	speaker's attribute frequency vector

All classifiers were trained, validated and tested using 10-fold cross validation, except for the case of the *Stars* corpus, in which we used 6-fold cross validation. The use of only six folds in the case of the *Stars* corpus reflects the number of descriptions available for each speaker in that corpus, which is kept balanced within each fold.

In order to optimize the C and γ parameter values of the SVM radial basis function kernel, grid search was performed. In every step of cross validation, one fold was left out as the test fold, and grid search was performed within the remaining folds, guided by the cross validation method as the performance metric on the training data. In the search, multiple SVM classifiers were attempted using different value combinations for C (1, 10, 100 and 1000) and γ (0.1, 0.01, 0.001, 0.0001). The winning parameter combinations were applied to train the SVM classifiers in all the remaining folds so as to make predictions in the test fold.

As in [5], we used Python *Scikit-learn* software for the actual SVM implementation. For the multi-class prediction of the relation attributes, we followed a 'one-against-one' approach [8].

4 Evaluation

Our work was evaluated against four corpora of definite descriptions: *GRE3D3* [13], *GRE3D7* [15], *Stars* [12] and *Stars2* [7]. In all cases, our goal was to compare the two models - with ($+SP$) and without ($-SP$) speaker-related features - as described in the previous section.

F-Score and Area under the ROC Curve (AUC) values for *GRE3D3/7* and *Stars/Stars2* individual classifiers are presented in Tables 3 and 4. Generally speaking, classifiers that took speakers' preferences into account outperformed those that did not in all corpora. The exceptions (*vpos* and *hpos* in *Stars* and *lm2_size* in *Stars2*) were due to data sparsity.

Each set of classifiers makes a REG algorithm as described in Section 3. Evaluation of the REG task was carried out by comparing the descriptions produced by each algorithm with the reference description found in each corpus. The overall precision of each algorithm was computed by measuring Accuracy (i.e., the number of exact matches between System and Reference description pairs). The degree of overlap between each System-Reference description pair was measured by Dice [3] scores. Results are summarized in Table 5.

Table 3. GRE3D3 and GRE3D7 classifier results

	GRE3D3				GRE3D7			
	-SP		+SP		-SP		+SP	
Classifier	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC
tg_type	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
tg_colour	0.88	0.16	0.95	0.94	0.99	0.35	0.99	0.85
tg_size	0.88	0.73	0.94	0.98	0.74	0.67	0.88	0.92
tg_location	0.00	0.15	0.30	0.74	0.00	0.23	0.00	0.75
lm_type	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
lm_colour	0.82	0.26	0.91	0.89	0.93	0.34	0.93	0.79
lm_size	0.73	0.58	0.77	0.90	0.66	0.52	0.84	0.87
lm_location	0.75	0.70	0.59	0.92	0.00	0.32	0.00	0.80
relation	0.15	0.39	0.87	0.96	0.00	0.50	0.75	0.91

Table 4. Stars and Stars2 classifier results

	Stars				Stars2			
	-SP		+SP		-SP		+SP	
Classifier	F_1	AUC	F_1	AUC	F_1	AUC	F_1	AUC
tg_type	1.00	1.00	1.00	1.00	0.99	0.27	1.00	0.86
tg_colour	0.00	0.27	0.69	0.87	0.71	0.80	0.80	0.94
tg_size	-	-	-	-	0.02	0.79	0.07	0.96
tg_hpos	0.00	0.23	0.47	0.79	0.00	0.00	0.00	0.00
tg_vpos	0.00	0.23	0.47	0.78	0.00	0.00	0.00	0.00
lm_type	0.99	0.31	1.00	0.75	0.99	0.15	0.99	0.43
lm_colour	0.58	0.43	0.79	0.81	0.84	0.84	0.88	0.95
lm_size	-	-	-	-	0.78	0.92	0.85	0.96
lm_hpos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lm_vpos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lm2_type	0.99	0.14	0.98	0.19	1.00	1.00	1.00	1.00
lm2_colour	0.20	0.30	0.68	0.79	0.00	0.36	0.35	0.88
lm2_size	-	-	-	-	0.00	0.00	0.00	0.00
lm2_hpos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
lm2_vpos	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
relation	0.98	0.22	0.98	0.66	0.90	0.85	0.95	0.97
lm_relation	0.00	0.34	0.73	0.89	0.00	0.40	0.43	0.70

Table 5. REG results with (+SP) and without (-SP) speaker-related features

	GRE3D3		GRE3D7		Stars		Stars2	
Alg.	Dice	Acc.	Dice	Acc.	Dice	Acc.	Dice	Acc.
-SP	0.78	0.46	0.88	0.61	0.72	0.18	0.66	0.30
+SP	0.92	0.74	0.94	0.77	0.73	0.29	0.76	0.36

We applied Wilcoxon's Signed-rank test over Dice scores, and the Chi-squared test over accuracy scores. Differences between the two algorithms are significant as summarized in Table 6.

Table 6. -SP and +SP results comparison

	Dice		Accuracy	
Corpus	W	p	χ^2	p
GRE3D3	1974.0	< .0001	245.75	< .0001
GRE3D7	261574.0	< .0001	475.27	< .0001
Stars	29927.0	< .8165	24.28	< .0001
Stars2	118278.0	< .0001	19.27	< .0001

Results show that taking speakers' preferences into account significantly increases both Dice and accuracy scores in all corpora, with the only exception of Dice scores on *Stars* data. This generally confirms our main research hypothesis.

The present results for *GRE3D3* are superior to those in [14], with reported 0.58 accuracy when using unpruned trees with 10-fold cross-validation, and also in [16], with reported Dice=0.85 and accuracy=0.60 for a 'longest first' selection strategy. Our results for *GRE3D7* are also superior to those in [14], with reported 0.67 accuracy when using pruned trees with 10-fold cross-validation.

Regarding *Stars* data, the present results are also superior to those obtained in [12], with reported Dice=0.61 and accuracy=0.11 when combining the Incremental algorithm with a decision-tree model of reference underspecification.

5 Final Remarks

This paper discussed the generation of referring expressions using SVM classifiers with and without speaker-related features. Results in four REG corpora suggest that the model that takes speakers' preferences into account outperforms the model that does not. Moreover, present results on *GRE3D3*, *GRE3D7* and *Stars* data are all superior to previous work in the field.

Speakers however do not stick to a single reference strategy in all situations, and may in fact produce different descriptions on different occasions even when the context remains unchanged. As future work, we intend to model this kind of non-determinism to account for both between-speakers and within-speakers variation in REG.

Acknowledgments. The authors acknowledge support by CAPES and FAPESP.

References

1. Bohnet, B.: The fingerprint of human referring expressions and their surface realization with graph transducers. In: INLG-2008. pp. 207–210. Stroudsburg, USA (2008)

2. Dale, R., Reiter, E.: Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science 19(2), 233–263 (1995)
3. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology 26(3), 297–302 (1945)
4. Fabbrizio, G.D., Stent, A.J., Bangalore, S.: Trainable speaker-based referring expression generation. In: Proceedings of the Twelfth Conference on Computational Natural Language Learning. pp. 151–158. CoNLL '08, Stroudsburg, PA, USA (2008), http://dl.acm.org/citation.cfm?id=1596324.1596350
5. Ferreira, T.C., Paraboni, I.: Classification-based referring expression generation. LNCS 8403, 481–491 (2014)
6. Gupta, S., Stent, A.J.: Automatic evaluation of referring expression generation using corpora. In: Proceedings of the 1st Workshop on Using Corpora in Natural Language Generation (UCNLG). pp. 1–6. Birmingham (2005)
7. Iacovelli, D., Galindo, M.R., Paraboni, I.: Lausanne: a framework for collaborative online NLP experiments. In: 11th International Conference on Computational Processing of Portuguese (PROPOR-2014). p. (to appear) (2014)
8. Knerr, S., Personnaz, L., Dreyfus, G.: Single-layer learning revisited: a stepwise procedure for building and training a neural network. In: Soulié, F., Hérault, J. (eds.) Neurocomputing, NATO ASI Series, vol. 68, pp. 41–50. Springer (1990)
9. Krahmer, E., van Deemter, K.: Computational generation of referring expressions: A survey. Computational Linguistics 38(1), 173–218 (2012)
10. de Lucena, D.J., Pereira, D.B., Paraboni, I.: From semantic properties to surface text: The generation of domain object descriptions. Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial 14(45), 48–58 (2010)
11. Pereira, D.B., Paraboni, I.: Statistical surface realisation of portuguese referring expressions. LNAI 5221, 383–392 (2008)
12. Teixeira, C.V.M., Paraboni, I., da Silva, A.S.R., Yamasaki, A.K.: Generating relational descriptions involving mutual disambiguation. LNCS 8403, 492–502 (2014)
13. Viethen, J., Dale, R.: The use of spatial relations in referring expression generation. In: INLG-2008. pp. 59–67. Stroudsburg, USA (2008)
14. Viethen, J., Dale, R.: Speaker-dependent variation in content selection for referring expression generation. In: Proceedings of the Australasian Language Technology Association Workshop 2010. pp. 81–89. Melbourne, Australia (December 2010)
15. Viethen, J., Dale, R.: GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In: Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop. pp. 12–22. Edinburgh, Scotland (July 2011)
16. Viethen, J., Mitchell, M., Krahmer, E.: Graphs and spatial relations in the generation of referring expressions. In: EACL-2013. pp. 72–81. Sofia (2013)