Improving the generation of personalised descriptions

Thiago Castro Ferreira
Tilburg center for Cognition and Communication (TiCC)
Tilburg University, The Netherlands
tcastrof@tilburguniversity.edu

Ivandré Paraboni
School of Arts, Sciences and Humanities
University of São Paulo, Brazil
ivandre@usp.br

Abstract

Referring expression generation (REG) models that use speaker-dependent information require a considerable amount of training data produced by every individual speaker, or may otherwise perform poorly. In this work we propose a simple personalised method for this task, in which speakers are grouped into profiles according to their referential behaviour. Intrinsic evaluation shows that the use of speaker's profiles generally outperforms the personalised method found in previous work.

1 Introduction

In natural language generation systems, referring expression generation (REG) is the microplanning task responsible for generating references of discourse entities (Krahmer and van Deemter, 2012). Choice of referential form (Ferreira et al., 2016), i.e., deciding whether a reference should be a proper name ('Ayrton Senna'), a pronoun ('He') or a description ('The racing driver'), is the first decision to be made in this task.

Albeit notable studies on pronominalisation (Callaway and Lester, 2002) and proper name generation (Ferreira et al., 2017), research on REG has largely focused on the generation of descriptions or, more specifically, on content selection. For instance, in the previous example, Ayrton Senna's *occupation* is the content selected to describe him. This work focuses on this kind of content selection task, hereby called REG for brevity.

Existing work in computational REG and related fields have identified a wide range of factors that may drive content selection. To a considerable extent, however, content selection is known to be influenced by human variation (Viethen and Dale, 2010). In other words, under identical circumstances (i.e., in the same referential context), different speakers will often produce different descriptions, and a single entity may be described by different speakers as 'the racing driver', 'the McLaren pilot', etc.

Existing REG algorithms as in Bohnet (2008) and Ferreira and Paraboni (2014) usually pay regard to human variation by computing personalised features from a training set of descriptions produced by each speaker. This highly personalised training method may of course be considered an ideal account of human variation but, in practice, will only be effective if every speaker in the domain is represented by a sufficiently large number of training instances.

As means to improve REG results when the amount of training data is limited, in this work we propose a simple training method for speaker-dependent REG in which training referring expressions are grouped into profiles according to the speaker's referential behaviour. The method relies on the observation that speakers tend to be consistent in their choices of referential overspecification, and it is shown to outperform the use of personalised information.

2 Related Work

Existing methods for speaker-dependent REG generally consist of computing the relevant features for each speaker. In what follows we summarise a number of studies that follow this method. In Bohnet (2008), the Incremental algorithm (Dale and Reiter,

Proceedings of The 10th International Natural Language Generation conference, pages 233–237,
Santiago de Compostela, Spain, September 4-7 2017. ©2017 Association for Computational Linguistics

1995) and a number of extensions of the Full Brevity algorithm (Dale, 1989) are evaluated on a corpus of furniture items and famous mathematicians (TUNA) (Gatt et al., 2007). In the case of the Incremental algorithm, human variation is accounted for by computing individual preference lists based on the attribute frequency of each speaker as observed in the training data. In the case of Full Brevity, all possible descriptions for a given referent are computed, and the description that most closely resembles those produced by the speaker is selected using a nearest neighbour approach.

The work in Viethen and Dale (2010) makes use of decision-tree induction to predict content patterns (i.e., full attribute sets representing actual referring expressions) to describe geometric objects on Google SketchUp scenes (GRE3D3/7 corpus) (Dale and Viethen, 2009; Viethen and Dale, 2011). Human variation is accounted for by modelling speaker identifiers as machine learning features.

Finally, the work in Ferreira and Paraboni (2014) presents a SVM-based approach to speaker-dependent REG tested also on the description of geometric objects (GRE3D3/7 and Stars/Stars2 (Teixeira et al., 2014; Paraboni et al., 2016) corpora). Once again, human variation is accounted for by computing individual preference lists from the subset of descriptions produced by each speaker.

3 Current work

In all the studies discussed in the previous section, personalised REG outperforms standard algorithms on domains in which a sufficient large number of training instances (i.e., referring expressions) is available for every speaker under consideration. However, the number of available instances per speaker tends to be small even in purpose-built REG corpora. For instance, there are only about 7 descriptions per speaker in the TUNA (singular) domain (Gatt et al., 2007), and 10-16 descriptions per speaker in GRE3D3/7 (Dale and Viethen, 2009; Viethen and Dale, 2011) and Stars/Stars2 (Teixeira et al., 2014; Paraboni et al., 2016).

To improve REG results in these situations, in what follows we consider a grouping personalised method that relies on psycholinguistic studies on referential overspecification (Koolen et al., 2011).

3.1 Basic REG model

We designed a REG experiment that makes use of a speaker-dependent REG model adapted from Ferreira and Paraboni (2014) as follows. Given a set D of domain objects, a set A of referential attributes, a set R of spatial relations between object pairs, and a target object $t \in D$ to be identified, content selection is implemented with the aid of a set of classifiers $C_{atom} = \{c^{(1)}, c^{(2)}, ..., c^{(|A|)}\}$, in which $c^{(i)} \in C_{atom}$ predicts whether $a^{(i)} \in A$ should be selected or not, and a multi-class classifier C_{rel} predicts the kind of relation ($r \in R$) that may hold between the target t and the nearest landmark lm. R includes the special *no-relation* property to denote situations in which no relation between a certain object pair is predicted. When a relation to a landmark object lm exists, we also consider a set of classifiers $C^{lm}_{atom} = \{c^{(1)}, c^{(2)}, ..., c^{(|A|)}\}$ to describe lm.

Part of the input to the classifiers consists of feature vectors extracted from the referential context. These features - hereby called context features - are based on the ones proposed in Viethen and Dale (2010), and are intended to model target and landmark properties (if any), and similarities between objects. More specifically, context features represent the size of the target and its nearest landmark, the relation (horizontal or vertical) between the two objects, and the number of distractors that share a certain property (e.g., *type*, *colour* etc.)

In order to model human variation, we also consider two kinds of speaker-dependent feature: those that model personal information about the speakers, and those that model their content selection preferences. Speaker's personal features consist of a unique speaker identifier as in Viethen and Dale (2010), gender and age bracket. Speaker's preferences consist of lists of preferred attributes for reference to target and landmark objects sorted by frequency. Attributes and relations of the main target t and nearby landmark lm are combined to form a description L according to Algorithm 1.

The input to the algorithm is a target t and a domain D. The algorithm also makes use of a history list H to prevent self-reference (e.g., 'the ball next to a box that is next to a ball that...') and the initially empty list L representing the output description (to be built recursively).

	TUNA-f		TUNA-p		GRE3D3		GRE3D7		Stars		Stars2		Overall	
Method	Dice	Acc.	Dice	Acc.	Dice	Acc.	Dice	Acc.	Dice	Acc.	Dice	Acc.	Dice	Acc.
Speaker	0.85	0.41	0.71	0.24	0.88	0.61	0.92	0.72	0.75	**0.39**	0.70	0.31	0.87	0.60
Profile	0.85	0.43	**0.78**	**0.35**	**0.93**	**0.74**	**0.94**	**0.77**	0.73	0.32	**0.78**	**0.40**	**0.90**	**0.66**

Table 1: Content selection results

Algorithm 1: Classification-based REG

```
Algorithm getDescription(t, L, D, H)
    $L[t] \leftarrow \{\}$
    $H \leftarrow H \cup t$
    $level \leftarrow |H|$
    $Pr_{atom} \leftarrow getPredictions(level)$
    $Pr_{rel} \leftarrow getRelationPrediction(level)$
    for $A_i \in Pr_{atom}$ do
        if $Pr_{atom}[A_i] == 1$ then
            $L[t] \leftarrow L[t] \cup \langle A_i, value(t, A_i) \rangle$
    if $Pr_{rel} \neq no\text{-}relation$ then
        $lm \leftarrow value(t, Pr_{rel})$
        if $lm \neq null$ and $lm \notin H$ then
            $L[t] \leftarrow L[t] \cup \langle rel, lm \rangle$
            $L \leftarrow getDescription(lm, L, D, H)$
    return $L$
```

An auxiliary function $level$ is assumed to return 1 when t corresponds to the main target, 2 when t corresponds to the first landmark object, and so on. This information is taken into account to invoke the appropriate set of classifiers, which are implemented by the auxiliary functions $getPredictions$ and $getRelationPrediction$. The former is assumed to invoke the set of binary classifiers for every attribute of t, and the latter invokes the multivalue prediction for the $relation$ class.

Content selection is performed by selecting all atomic attributes of the target t that were predicted by the corresponding binary classifiers. When a relation between t and its nearest distractor lm is predicted, the relation is included in L and the algorithm is called recursively to describe lm as well.

3.2 Personalised method

As an alternative to standard speaker-dependent REG (which relies on a set of descriptions produced by each speaker as in, e.g., Bohnet (2008)), we propose a personalised method based on the simple observation - made by Viethen and Dale (2010) and others - that some speakers follow a consistent pattern in reference production, whereas others do not.

In the present method - hereby called *Profile* - speakers are divided into three simple categories: those that always produced overspecified descriptions, those that always produced minimally distinguishing descriptions, and those that do not follow a consistent pattern. Knowing in advance the category of a particular speaker, the REG model will be trained on the descriptions produced by that category only. This will effectively allow us to use more training data than in standard personalised methods.

4 Evaluation

Data Six REG datasets: TUNA-Furniture and TUNA-People (Gatt et al., 2007) (in both cases, only descriptions to single objects were considered), GRE3D3 (Dale and Viethen, 2009), GRE3D7 (Viethen and Dale, 2011), Stars (Teixeira et al., 2014) and Stars2 (Paraboni et al., 2016).

Models As in Ferreira and Paraboni (2014), all classifiers were built using Support Vector Machines (SVMs) with a Gaussian Kernel. For the relation prediction, we use an "one-against-one" multi-class method. All models were evaluated using cross-validation with a balanced number of referring expressions per participant within each fold. For TUNA and Stars, descriptions were divided into six folds each. For GRE3D3/7 and Stars2, descriptions were divided into ten folds each. Grid-search was used to obtain an optimal model setting by testing values for the SVM C parameter (1, 10, 100 and 1000) and the Gaussian kernel γ (1, 0.1, 0.01, and 0.001) in a validation set before the test step.

Baseline We make use of a baseline method called *Speaker*. In this method, classifiers are trained on the set of referring expressions produced by each individual speaker.

Metrics We measured Dice coefficients (Dice, 1945) to assess the similarity between each description generated by the model and the corpus description. We also computed the overall REG Accuracy

Method	TUNA-f	TUNA-p	GRE3D3	GRE3D7	Stars	Stars2	Overall
Speaker	0.75	0.70	0.54	0.80	0.70	0.65	0.75
Profile	0.78	0.78	0.61	0.82	0.68	0.78	0.79

Table 2: Reference type classification for each corpus

by counting the number of exact matches between each description pair.

5 Results

Table 1 presents the results of the REG model using the *Speaker* and *Profile* personalised methods on each of the test domains. Overall results suggest that *Profile* outperforms *Speaker* both in terms of Dice (Wilcoxon W=3188296.5, $p<0.01$) and Accuracy (Chi-Square $\chi^2 = 104.28$, $p<0.01$) scores.

Regarding the results in individual domains, we notice that *Profile* outperforms *Speaker* in terms of Dice scores in the case of TUNA-People, GRE3D3, GRE3D7 and Stars2. A pairwise comparison shows that these differences are significant at $p<.01$. In the case of TUNA-Furniture and Stars the difference was not significant. *Profile* also outperforms *Speaker* in terms of Accuracy in TUNA-People, GRE3D3, GRE3D7 and Stars2, with pairwise comparisons significant at $p<0.01$. In the case of TUNA-Furniture, the difference was not significant, and in the case of Stars a significant effect in the opposite direction was observed ($\chi^2 = 9.38$, $p<0.01$).

Finally, Table 2 shows how often the *Speaker* and *Profile* methods were able to reproduce the level of referential specification found in the corpus, that is, how often each method correctly produced underspecified, overspecified and minimally distinguishing descriptions. Results show that predictions made by the *Profile* method generally outperform those made by the *Speaker* method, the exception being the case of the Stars corpus.

6 Discussion

This paper presented a machine-learning approach to REG that takes speaker-dependent information into account by making use of a personalised method to circumnavigates the issue of data sparsity. By grouping speakers according to a simple model of referential overspecification, we were arguably able to sketch a more general approach to speaker-dependent REG that was shown to outperform the standard use of individual speaker's information proposed in previous work.

Since using more training data - as we did by considering groups of similar speakers - improved results, we may of course argue that by simply training our REG models on the data provided by *all* speakers may improve results even further. Although we presently do not seek to validate this claim, there is plenty of evidence to suggest that this would not be the case. Studies such as in Bohnet (2008) have consistently shown that using individual training datasets for each speaker outperforms speaker-independent REG and, in particular, the work in Ferreira and Paraboni (2014) has shown that the current SVM model produces best results when trained on personalised datasets.

7 Future Work

The low availability of training data is not the only challenge to be dealt with in speaker-dependent REG. We notice that there is also the related issue of domain complexity. Existing REG models usually assume the existence of a pre-defined knowledge base of entities and their properties (Dale and Haddock, 1991; Dale and Reiter, 1995) or, as in the present case, take into account an overly simplified domain that restricts content selection. As a result, the variation in the output descriptions is limited by the knowledge base.

In future, the issue may be addressed by using the *semantic web* as the input to the REG model. This strategy, which has been shown succeed in the generation of proper names (Ferreira et al., 2017), may provide more information about the entities and their relations, and allow the generation of descriptions with greater variation (and possibly closer to the descriptions produced by any particular individual).

Acknowledgements

This work has been supported by the National Council of Scientific and Technological Development from Brazil (CNPq) and FAPESP 2016/14223-0.

References

B. Bohnet. 2008. The fingerprint of human referring expressions and their surface realization with graph transducers. In *Fifth International Natural Language Generation Conference*, pages 207–210, Stroudsburg, PA, USA.

C. B. Callaway and J. L. Lester. 2002. Pronominalization in generated discourse and dialogue. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 88–95, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

R. Dale and N. J. Haddock. 1991. Content determination in the generation of referring expressions. *Computational Intelligence*, 7(4):252–265.

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

R. Dale and J. Viethen. 2009. Referring expression generation through attribute-based heuristics. In *Proceedings of ENLG-2009*, pages 58–65.

R. Dale. 1989. Cooking up referring expressions. In *Proc. ACL-1989*, pages 68–75, Stroudsburg, USA.

L. R. Dice. 1945. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302.

T. C. Ferreira and I. Paraboni. 2014. Referring expression generation: taking speakers' preferences into account. *Lecture Notes in Artificial Intelligence*, 8655:539–546.

T. C. Ferreira, E. Krahmer, and S. Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany, August. Association for Computational Linguistics.

T. C. Ferreira, E. Krahmer, and S. Wubben. 2017. Generating flexible proper name references in text: Data, models and evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 655–664, Valencia, Spain, April. Association for Computational Linguistics.

A. Gatt, I. van der Sluis, and K. van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of ENLG-07*.

R. Koolen, A. Gatt, M. Goudbeek, and E. Krahmer. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics*, 43(13):3231–3250.

E. Krahmer and K. van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

I. Paraboni, M. Galindo, and D. Iacovelli. 2016. Stars2: a corpus of object descriptions in a visual domain. *Language Resources and Evaluation*.

C. V. M. Teixeira, I. Paraboni, A. S. R. da Silva, and A. K. Yamasaki. 2014. Generating relational descriptions involving mutual disambiguation. *LNCS*, 8403:492–502.

J. Viethen and R. Dale. 2010. Speaker-dependent variation in content selection for referring expression generation. In *Australasian Language Technology Association Workshop 2010*, pages 81–89, Melbourne, Australia.

J. Viethen and R. Dale. 2011. GRE3D7: A corpus of distinguishing descriptions for objects in visual scenes. In *Proceedings of UCNLG+Eval-2011*, pages 12–22.