Generating overspecified referring expressions: the role of discrimination

Ivandré Paraboni, Michelle Reis Galindo, Douglas Iacovelli
School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
Av. Arlindo Bettio, 1000 - São Paulo, Brazil
{ivandre,michelle.galindo,douglas.iacovelli}@usp.br

Abstract

We present an experiment to compare a standard, minimally distinguishing algorithm for the generation of relational referring expressions with two alternatives that produce overspecified descriptions. The experiment shows that discrimination - which normally plays a major role in the disambiguation task - is also a major influence in referential overspecification, even though disambiguation is in principle not relevant.

1 Introduction

In Natural Language Generation (NLG) systems, Referring Expression Generation (REG) is the computational task of providing natural language descriptions of domain entities (Levelt, 1989; Dale and Reiter, 1995), as in 'the second street on the left', 'the money that I found in the kitchen' etc. In this paper we will focus on the issue of content selection of *relational* descriptions, that is, those in which the intended target is described *via* another object, hereby called a *landmark*. Consider the example of context in Fig. 1.

Figure 1: A simple visual context. All objects are grey except for $obj5$, which is red.

Let us consider the goal of uniquely identifying the target $obj1$ in the context in Fig.1. Since the target shares most atomic properties (e.g., type, colour and size) with other distractor objects in the context (and particularly so with respect to $obj4$), using a relational property (*near-obj2*) may help prevent ambiguity. The following (a)-(c) are examples of descriptions of this kind produced from the above context.

(a)The cone near the box
(b)The cone near the *grey* box
(c)The cone near the *small* box

As in example (a), existing REG algorithms will usually pay regard to the Gricean maxim of quantity (Grice, 1975), and avoid the inclusion of properties that are not strictly required for disambiguation. In the case of relational reference, this means that both target and landmark portions of the description may be left *underspecified*, and uniqueness will follow from the fact that they mutually disambiguate each other (Teixeira et al., 2014). In other words, example (a) may be considered felicitous even though both 'cone' and 'box' are ambiguous if interpreted independently.

Minimally distinguishing descriptions as in (a) are the standard output of many REG algorithms that handle relational descriptions as in (Dale and Haddock, 1991; Krahmer and Theune, 2002; Krahmer et al., 2003). Human speakers, on the other hand, are largely redundant (Engelhardt et al., 2006; Arts et al., 2011; Koolen et al., 2011; Engelhardt et al., 2011), and will often produce so-called *overspecified* descriptions as in (b-c) above.

In this paper we will focus on the issue of generating overspecified relational descriptions as in examples (b-c), discussing which properties should be selected by a REG algorithm assuming that the decision to overspecify has already been made. More specifically, we will discuss whether the algorithm should include colour as in (b), size as in (c), or other alternatives, and we will assess the impact of a referential overspecification strategy that favours highly discriminatory properties over preferences that are well-established in the literature. Although this may in principle seem as a narrow research topic, the generation of relational descriptions is still subject of considerable debate in the field (e.g., (Viethen and Dale, 2011) and

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 76–82,
Beijing, China, July 26-31, 2015.

the issue of landmark under/full-specification has a number of known consequences for referential identification (e.g., (Paraboni and van Deemter, 2014)).

2 Related work

2.1 Relational REG

One of the first REG algorithms to take relations into account is the work in (Dale and Haddock, 1991), which generates descriptions that may include relational properties only as a last resort, that is, only when it is not possible to obtain a uniquely identifying descriptions by making use of a set of atomic properties. The algorithm prevents circularity (e.g., 'the cup on the table that supports a cup that...') and avoids the inclusion of redundant properties with the aid of consistency networks. As a result, the algorithm favours the generation of minimally distinguishing relational descriptions as example (a) in the previous section.

In the Graph algorithm described in (Krahmer et al., 2003), the referential context is modelled as a labelled directed graph with vertices representing domain entities and edges representing properties that can be either relational (when connecting two entities) or atomic (when forming self-loops). The task of obtaining a uniquely identifying description is implemented as a subgraph construction problem driven by domain-dependent cost functions associated with the decisions made by the algorithm. The work in (Krahmer et al., 2003) does not make specific assumptions about the actual attribute selection policy, and by varying the cost functions it is possible to implement a wide range of referential strategies. The use of the algorithm for the generation of relational descriptions is discussed in (Viethen et al., 2013).

The work in (Paraboni et al., 2006) discusses the issue of ease of search by focussing on the particular case of relational description in hierarchically-ordered domains (e.g., books divided into sections and subsections etc.) Descriptions that may arguably make search difficult, as in 'the section that contains a picture' are prevented by producing fully-specified descriptions of each individual object (i.e., picture, section etc.). As in (Dale and Haddock, 1991), atomic properties are always attempted first, and each target (e.g., a subsection) holds only one relation (e.g., to its parent section). Descriptions of this kind are similar to the examples (b-c) in the previous section. However, hierarchical structures are highly specialised domains, and it is less clear to which extent these findings are applicable to more general situations of reference as in, e.g., spatial domains (Byron et al., 2007; dos Santos Silva and Paraboni, 2015).

2.2 Referential overspecification

Assuming that we would like to add a redundant property to overspecify a certain description, which property should be selected? Research on REG, cognitive sciences and related fields has investigated a number of factors that may play a role in referential overspecification. First of all, it has been widely observed that some properties are simply preferred to others. This seems to be the case, for instance, of the colour attribute. Colour is ubiquitously found in both redundant and non-redundant use (Pechmann, 1989), and empirical evidence suggests that colour is overspecified more frequently than size (Belke and Meyer, 2002).

The inherent preference for colour has however been recently challenged. The work in (van Gompel et al., 2014), for instance, points out that when perceptual salience is manipulated so that a high contrast between target and distractors is observed, the size attribute may be preferred to colour. In other words, a highly preferred property may not necessarily match the choices made by human speakers when producing overspecified descriptions. Results along these lines are also reported in (Tarenskeen et al., 2014).

Redundant and non-redundant uses of colour (and possibly other preferred properties) may also be influenced by the difficulty in encoding visual properties. In (Viethen et al., 2012), for instance, it is argued that the colour property is more likely to be selected when it is maximally different from the other colours in the context. For instance, a red object is more likely to be described as 'red' when none of the distractors is red, and less so when a modifier (e.g., 'light red') would be required for disambiguation.

Closer to our present discussion, we notice that the issue of discrimination as proposed in (Olson, 1970) has been considered by most REG algorithms to date (e.g., (Dale and Reiter, 1995; Krahmer and van Deemter, 2012)), and it has even motivated a number of greedy or minimally distinguishing REG strategies (Gardent, 2002; Dale, 2002; Areces et al., 2011). Interestingly, the work

in (Gatt et al., 2013) has suggested that small differences in discriminatory power do not seem to influence content selection, but large differences do, a notion that has been applied to the design of REG algorithms on at least two occasions: in (de Lucena et al., 2010) properties are selected in order of preference regardless of their discriminatory power and, if necessary, an additional, highly discriminatory property is included; in (van Gompel et al., 2012), a fully distinguishing property is attempted first and, if necessary for disambiguation, further properties are considered based on both preference and discrimination.

Discrimination clearly plays a major role in the disambiguation task, but it less clear whether it is still relevant when disambiguation is not an issue, that is, in the case of referential overspecification. The present work is an attempt to shed light on this particular issue.

3 Current work

Following (Pechmann, 1989) and others, we may assume that colour should be generally (or perhaps always) preferred to size. Moreover, as in (Kelleher and Kruijff, 2006), we may follow the principle of minimal effort (Clark and Wilkes-Gibbs, 1986) and assume that atomic properties such as colour or size should be preferred to relations that lead to more complex descriptions. In our current work, however, we will argue that neither needs to be the case: under the right circumstances, a wide range of properties - colour, size and even spatial relations - may be overspecified depending on their *discriminatory power* alone. Thus, it may be the case that size is preferred to colour (unlike, e.g., (Pechmann, 1989)), and that longer, relational descriptions are preferred to shorter ones (unlike, e.g., (Kelleher and Kruijff, 2006)).

The possible preference for highly discriminatory properties in referential overspecification is easily illustrated by the examples in the introduction section. Following (Pechmann, 1989), one might assume that, if a speaker decides to overspecify the landmark portion of description (a), she may add the colour attribute, as in (b). This strategy, however, turns out to be far less common in language use if a more discriminatory property is available, as in the example. More specifically, the availability of a highly discriminatory landmark property (*size-small*) makes (c) much more likely than (b). This observation gives rise to the following research hypothesis:

> *h1: Given the goal of overspecifying a relational description by using an additional landmark property p, p should correspond to the most discriminatory property available in the context.*

The idea that speakers may take discriminatory power into account when referring is of course not novel. What is less obvious, however, is that discrimination may also play a significant role in situations that do not involve ambiguity, as in the above examples. To illustrate this, let us consider a basic REG algorithm - hereby called *Baseline* - consisting of a relational implementation of an Incremental-like algorithm as proposed in (Dale and Reiter, 1995).

Given the goal of producing a uniquely identifying description L of a target object r, the *Baseline* algorithm works as follows: first, an atomic description is attempted by examining a list of preferred attributes P and by selecting those that help disambiguate the reference, as in the standard Incremental approach (Dale and Reiter, 1995). If the description is uniquely identifying, the algorithm terminates. If not, a relational property relating r to a landmark object o is included in L, and the algorithm is called recursively to describe o using an atomic description if possible.

Since *Baseline* terminates as soon as a uniquely identifying description is obtained, the landmark description will be usually left underspecified as in example (a) in Section 1. This behaviour is consistent with existing relational REG algorithms (e.g., (Dale and Haddock, 1991; Krahmer et al., 2003)).

Using the *Baseline* descriptions as a starting point, however, we may decide to fully-specify the landmark description (e.g., in order to facilitate search, as in (Paraboni and van Deemter, 2014)) by selecting an additional property p from the remainder P list, hereby called P_0.

There are of course many ways of defining p. In corpus-based REG, for instance, a plausible strategy would be to assume that the definition of p is domain-dependent, and simply select the most frequent (but still discriminatory) property in P_0 as seen in training data. We will call this variation the *Most Frequent* overspecification strategy.

Choosing the most frequent property p may lead to descriptions that closely resemble those observed in the data. However, we predict that

the availability of a highly discriminatory property may change this preference. To illustrate this, we will also consider a *Proposal* strategy in which p is taken to be the most discriminatory property available in P_0. In case of a tie, the most frequent property that appears in P_0 is selected. If P_0 does not contain any discriminatory properties, none will be selected and the landmark description will remain underspecified as in the standard *Baseline* approach.

The context in the previous Fig.1 and the accompanying examples (a-c) in Section 1 illustrate the expected output of each of the three algorithms under consideration. As in previous work on relational REG, the *Baseline* approach would produce the minimally distinguishing description (a); the *Most Frequent* strategy would overspecify the landmark portion of the description by adding the preferred property in the relevant domain (e.g., colour) as in (b); and the *Proposal* strategy would overspecify by adding the highly discriminatory property (in this particular example, size) as in (c).

The relation between the three algorithms and our research hypothesis $h1$ is straightforward. We would like to show that the predictions made by *Proposal* are more accurate than those made by *Baseline* and *Most Frequent*. An experiment to verify this claim is described in the next section.

4 Experiment

For evaluation purposes we will make use the Stars2 corpus of referring expressions[1]. Stars2 is an obvious choice for our experiment since these data convey visual scenes in which objects will usually have one highly discriminatory property available for reference. Moreover, descriptions in this domain may convey up to two relations (e.g., 'the cone next to the ball, near the cone'), which gives rise to multiple opportunities for referential overspecification.

In addition to this, we will also make use of the subset of relational descriptions available from the GRE3D3 (Dale and Viethen, 2009) and GRE3D7 (Viethen and Dale, 2011) corpora. Situations of reference in the GRE3D3/7 domain are in many ways simpler than those in Stars2 (i.e., by containing at most one possible relation in each scene, by not presenting any property whose discriminatory power is substantially higher than others etc.), but the comparison is still useful since GRE3D3/7 are among the very few annotated relational REG corpora made publicly available for research purposes, and which have been extensively used in previous work.

From the three domains - Stars2, GRE3D3 and GRE3D7 - we selected all instances of relational descriptions in which the landmark object was described by making use of the *type* attribute and exactly one additional property p. This amounts to three *Reference* sets containing 725 descriptions in total: 367 descriptions from Stars2, 114 from GRE3D3 and 244 from GRE3D7.

In the situations of reference available from these domains, the use of p is never necessary for disambiguation, and p will never be selected by a standard REG algorithm as the *Baseline* strategy described in the previous section. Thus, our goal is to investigate which overspecification strategy - *Proposal* or *Most Frequent*, cf. previous section - will select the correct p, and the corresponding impact of this decision on the overall results of each algorithm.

From the unused portion of each corpus, we estimate attribute frequencies to create the preference list P required by the algorithms. The following preference orders were obtained:

$P(Stars2)$ ={*type*, *colour*, *size*, *near*, *in-front-of*, *right*, *left*, *below*, *above*, *behind*}

$P(GRE3D)$ ={*type*, *colour*, *size*, *above*, *in-front-of*, *hpos*, *vpos*, *near*, *right*, *left*}

In the case of the GRE3D3/7 corpora, we notice that not all attributes appear in both data sets. Moreover, the attributes *hpos* and *vpos* were computed from the existing *pos* attribute, which was originally intended to model both horizontal and vertical screen coordinates as a single property in (Dale and Viethen, 2009).

Each of the three REG strategies - *Baseline*, *Proposal* and *Most Frequent* - received as an input the 725 situations of reference represented in the *Reference* data and the corresponding P list for each domain. As a result, three sets of output descriptions were obtained, hereby called *System* sets.

Evaluation was carried out by comparing each *System* set to the corresponding *Reference* corpus descriptions and measuring *Dice* scores (Dice, 1945) and overall accuracy (that is, the number of exact matches between each *System-Reference* description pair).

[1]Some of the corpus features are described in (Ferreira and Paraboni, 2014)

Table 1: Results

Algorithm	*Baseline*				*Most frequent*				*Proposal*			
	Dice		Accuracy		Dice		Accuracy		Dice		Accuracy	
Dataset	mean	sdv	mean	sdv	mean	sdv	mean	sdv	mean	sdv	mean	sdv
Stars2	0.63	0.14	0.00	0.00	0.62	0.18	0.11	0.31	**0.76**	0.18	**0.27**	0.45
GRE3D3	0.81	0.06	0.00	0.00	0.87	0.10	0.25	0.43	**0.90**	0.09	**0.36**	0.48
GRE3D7	0.84	0.07	0.00	0.00	**0.92**	0.10	**0.47**	0.50	0.89	0.10	0.34	0.48
Overall	0.73	0.15	0.00	0.00	0.76	0.21	0.25	0.43	**0.82**	0.16	**0.31**	0.46

5 Results

Table 1 shows descriptive statistics for the evaluation of our three algorithms - *Baseline*, *Proposal* and *Most Frequent* - applied to each corpus - Stars2, GRE3D3 and GRE3D7. Best results are highlighted in boldface.

Following (Gatt and Belz, 2007) and many others, we compare *Dice* scores obtained by the three algorithms applied to the generation of the selected descriptions of each domain using *Wilcoxon's* signed-rank test. In the Overall evaluation, *Proposal* outperforms both alternatives. The difference is significant (W(338)=-34327, Z=-9.55, $p < 0.0001$). Highly discriminatory properties are indeed those that are normally selected by human speakers when they decide to overspecify a landmark description. This supports our research hypothesis $h1$.

Individual results are as follows. In the case of the Stars2 domain, *Proposal* outperforms both alternatives. The difference is significant (W(241)=-26639, Z=-12.29, $p < 0.0001$). In the case of GRE3D3, once again *Proposal* outperforms the alternatives. The difference is also significant (W(27)=-248, Z=-2.97, $p < 0.03$). Finally, in the case of GRE3D7, an effect in the opposition direction was observed, i.e., the *Most Frequent* algorithm outperforms the alternatives. The difference is significant (W(70)=1477, Z=4.32, $p < 0.0001$).

The differences across domains are explained by the proportion of highly discriminatory landmark properties in each corpus. In Stars2, the nearest landmark has at least one highly discriminatory property in all scenes involving relational reference. In GRE3D3, the nearest landmark has a highly discriminatory property in 80% of the scenes, and in GRE3D7 this is the case in only 50% of the scenes. Thus, given the opportunity, the use of a highly discriminatory property seems to be preferred. The absence of a property that 'stands out', by contrast, appears to make the choice among them a matter of preference, an observation that is consistent with the findings in (Gatt et al., 2013).

6 Final remarks

This paper has presented a practical REG experiment to illustrate the impact of discrimination on the generation of overspecified relational descriptions. The experiment shows that discrimination - which normally plays a major role in the disambiguation task - is also a considerable influence in referential overspecification, that is, even when discrimination is in principle not an issue. Our findings correlate with previous empirical work in the field, and show that discrimination may effectively trump the inherent preference for absolute properties and for those that are easier to realise in surface form. For instance, contrary to (Pechmann, 1989) and many others, speakers would generally prefer referring to size as in (b), despite evidence suggesting that colour is overspecified more frequently than size. Moreover, contrary to (Kelleher and Kruijff, 2006), speakers would also prefer referring to a spatial relation as in (c) even though the resulting descriptions turns out to be more complex.

We are aware that the present work has focussed on extreme situations in which a highly discriminatory property is available for overspecification. As future work, it is necessary to further this investigation by taking into account various degrees of discrimination. As suggested in (Gatt et al., 2013), the effect of discrimination may be perceived as a continuum, and in that case a practical REG algorithm should be able to make more complex decisions that those presently implemented.

Acknowledgements

This work has been supported by FAPESP and by the University of São Paulo.

References

C. Areces, S. Figueira, and D. Gorín. 2011. Using logic in the generation of referring expressions. In *Proceedings of the 6th International Conference on Logical Aspects of Computational Linguistics (LACL 2011)*, pages 17–32, Montpelier. Springer.

A. Arts, A. Maes, L. G. M. Noordman, and C. Jansen. 2011. Overspecification facilitates object identification. *Journal of Pragmatics*, 43(1):361–374.

E. Belke and A. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination. *European Journal of Cognitive Psychology*, 14(2):237–266.

D. Byron, A. Koller, J. Oberlander, L. Stoia, and K. Striegnitz. 2007. Generating instructions in virtual environments (GIVE): A challenge and evaluation testbed for NLG. In *Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*.

H. Clark and D. Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22:1–39.

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.

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.

Diego Jesus de Lucena, Ivandré Paraboni, and Daniel Bastos Pereira. 2010. From semantic properties to surface text: The generation of domain object descriptions. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial*, 14(45):48–58.

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, K. Baileyand, and 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.

Thiago Castro Ferreira and Ivandré Paraboni. 2014. Referring expression generation: taking speakers' preferences into account. *Lecture Notes in Artificial Intelligence*, 8655:539–546.

C. Gardent. 2002. Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 96–103.

Albert Gatt and Anja Belz. 2007. The attribute selection for GRE challenge: Overview and evaluation results. In *UCNLG+MT: Language Generation and Machine Translation*.

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.

H. P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics*, volume 3. New York: Academic Press.

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.

E. 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.

W. Levelt. 1989. *Speaking: From intention to articulation*. MIT press, Cambridge, Ma.

D. R. Olson. 1970. Language and thought: aspects of a cognitive theory of semantics. *Psychological Review*, 77(4):257–273.

Ivandré Paraboni and Kees van Deemter. 2014. Reference and the facilitation of search in spatial domains. *Language, Cognition and Neuroscience*, 29(8):1002–1017.

Ivandré Paraboni, Judith Masthoff, and Kees van Deemter. 2006. Overspecified reference in hierarchical domains: measuring the benefits for readers. In *Proc. of INLG-2006*, pages 55–62, Sydney.

T. Pechmann. 1989. Incremental speech production and referential overspecification. *Linguistics*, 27(1):98–110.

Sammie Tarenskeen, Mirjam Broersma, and Bart Geurts. 2014. Referential overspecification: Colour is not that special. In *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh, Scotland.

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*.

Roger van Gompel, Albert Gatt, Emiel Krahmer, and Kees Van Deemter. 2014. Testing computational models of reference generation as models of human language production: The case of size contrast. In *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh, Scotland.

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, Martijn Goudbeek, and Emiel Krahmer. 2012. The impact of colour difference and colour codability on reference production. In *Proceedings of CogSci-2012*, pages 1084–1098.

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.