The Costs of Redundancy in Referring Expressions (GRAPH)

Emiel Krahmer
Tilburg University
The Netherlands
e.j.krahmer@uvt.nl

Mariët Theune
University of Twente
The Netherlands
m.theune@utwente.nl

Jette Viethen
Macquarie University
Australia
jviethen@ics.mq.edu.au

Iris Hendrickx
University of Antwerp
Belgium
iris.hendrickx@ua.ac.be

Abstract

We describe a graph-based generation system, which participated in the Attribute Selection Task of REG 2008. Using a stochastic cost function (with certain properties for free), and trying attributes from cheapest to more expensive, the system achieves overall .76 DICE and .54 MASI scores on the development set.

1 Introduction

Referring Expression Generation (REG) is a key-task in NLG, and the topic of the REG 2008 Challenge.[1] In this context, referring expressions are understood as *distinguishing descriptions*: descriptions that uniquely characterize a target object in a visual scene (e.g., "the red sofa"), and do not apply to any of the other objects in the scene (the distractors). Generating such descriptions is usually assumed to be a two-step procedure: first, it has to be decided which attributes of the target suffice to characterize it uniquely, and then the selected set of attributes should be converted into natural language.

Here we focus on the first step (attribute selection), which corresponds to Task 1 (TUNA-AS) in the REG 2008 Challenge, and for this we use a version of the Graph-based REG algorithm of Krahmer et al. (2003). In this approach, a visual scene is represented as a directed labelled graph, where vertices represent the objects in the scene and edges their attributes. A key ingredient of the approach is that costs can be assigned to attributes; the generation of referring expressions can then be defined as a graph search problem, which outputs the cheapest distinguishing graph (if one exists) given a particular cost function.

A version of this algorithm was submitted for the ASGRE 2007[2] Challenge (Theune et al. 2007). For us, one of the most striking, general outcomes was the observed "trend for the mean DICE score obtained by a system to decrease as the proportion of minimal descriptions increases" (Belz and Gatt 2007).[3] Thus, while REG systems have a tendency to produce minimal descriptions, human speakers tend to include redundant properties in their descriptions, which is in line with recent findings in psycholinguistics on the production of referring expressions (e.g., Engelhardt et al. 2006).

In principle, the graph-based approach has the potential to deal with redundancy by allowing some attributes to have zero costs. Viethen et al. (2008), however, show that merely assigning zero costs to an attribute is not a sufficient condition for inclusion; if the search terminates before the free properties are tried, they will not be included. In other words: the order in which attributes are tried should be explicitly controlled as well. In the experiment we describe here, we consider both these factors and their interplay.

2 Method

We experimentally combine four cost functions and two search orders (Table 1). (1) **Simple** simply assigns each edge a 1-point cost. (2) **Stochastic** asso-

[1] See http://www.itri.brighton.ac.uk/research/reg08/.

[2] See http://www.csd.abdn.ac.uk/research/evaluation/.

[3] DICE (like MASI) is a measure for similarity between a predicted attribute set and a (human produced) reference set.

Table 2: Results for the 2008 development set. The GRAPH 4+B settings were submitted to the REG 2008 Challenge.

Graph System	Furniture				People				Overall			
	DICE	MASI	ACC	MIN	DICE	MASI	ACC	MIN	DICE	MASI	ACC	MIN
1+A	.61	.32	.12	.29	.59	.36	.24	.00	.60	.34	.18	.16
1+B	.61	.31	.12	.29	.66	.42	.24	.00	.63	.36	.18	.16
2+A	.71	.47	.31	.11	.66	.42	.24	.00	.69	.45	.28	.06
2+B	.69	.44	.28	.16	.66	.42	.24	.00	.68	.43	.26	.09
3+A	.80	.58	.45	.00	.68	.41	.19	.00	.74	.51	.33	.00
3+B	.80	.58	.45	.00	.72	.48	.28	.00	.76	.54	.37	.00
4+A	.80	.59	.48	.00	.59	.34	.18	.00	.70	.48	.34	.00
4+B	.80	.59	.48	.00	.72	.48	.28	.00	.76	.54	.39	.00
FP (2007)	.71	–	–	–	.67	–	–	–	.69	–	–	–

Table 1: Overview of cost functions and search orders.

Costs	Version
1	Simple
2	Stochastic
3	Free-stochastic
4	Free-naive
Orders	**Version**
A	Random
B	Cost-based

ciates each edge with a frequency-based cost, based on both the 2008 training and development sets (assuming that a larger data set allows for more accurate frequency estimates). (3) **Free-Stochastic** is like the previous cost function, except that highly frequent attributes are assigned 0 costs. For the Furniture domain, this applies to "colour"; for People to "hasBeard = 1" and "hasGlasses = 1." (4) **Free-Naive**, finally, reduces the relatively fine-grained costs of Free-Stochastic to three values (0 = free, 1 = cheap, 2 = expensive). In addition, we compare results for two property orderings: (A) Properties are tried in a **Random** order. (B) **Cost-based**, where properties are tried (in stochastic order) from cheapest to most expensive. Finally, since human speakers nearly always include the "type" property, we decided to simply always include it.

3 Results and Discussion

Table 2 summarizes the results. Notice first that the order in which attributes are tried in the search process matters; the B-systems nearly always outperform their A-counterparts. Second, assigning varying costs also helps; both 1-variants (**Simple** costs) perform worse than the systems building on **Stochastic** cost functions (2, 3 and 4). Third, adding free properties is also beneficial; the 3 and 4 variants clearly outperform the 1 and 2 variants. It is interesting to observe that the **Free-naive** cost function (4) performs equally well as the more principled **Free-stochastic** (3), but only in combination with the **Cost-based** order (B). To the extent that it is possible to compare the results, the submitted GRAPH 4+B outperforms our best 2007 variant (GRAPH FP in Table 2). This suggests that the interplay between property ordering and cost function is a flexible and efficient approach to attribute selection.

References

Belz, A. and A. Gatt 2007. The attribute selection for GRE challenge: Overview and evaluation results *Proceedings of UCNLG+MT* 75-83

Engelhardt, P., K. Bailey and F. Ferreira 2006. Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language, 54*, 554-573.

Krahmer, E., S. van Erk and A. Verleg 2003. Graph-based generation of referring expressions. *Computational Linguistics, 29(1)*, 5372.

Theune, M., P. Touset, J. Viethen, and E. Krahmer. 2007. Cost-based attribute selection for generating referring expressions (GRAPH-FP and GRAPH-SC). *Proceedings of the* ASGRE *Challenge 2007*, Copenhagen, Denmark

Viethen, J., R. Dale, E. Krahmer, M. Theune and P. Touset. 2008. Controlling redundancy in referring expressions. *Proceedings LREC 08*, Marrekech, Morroco.