COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 36 (2012) 837–841

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2012.01258.x

Is It That Difficult to Find a Good Preference Order for the Incremental Algorithm?

Emiel Krahmer,[a] Ruud Koolen,[a] Mariët Theune[b]

[a]*Tilburg Center for Cognition and Communication (TiCC),Tilburg University*
[b]*Human-Media Interaction, University of Twente*

Received 1 November 2011; received in revised form 13 February 2012; accepted 15 February 2012

Abstract

In a recent article published in this journal (van Deemter, Gatt, van der Sluis, & Power, 2012), the authors criticize the Incremental Algorithm (a well-known algorithm for the generation of referring expressions due to Dale & Reiter, 1995, also in this journal) because of its strong reliance on a predetermined, domain-dependent Preference Order. The authors argue that there are potentially many different Preference Orders that could be considered, while often no evidence is available to determine which is a good one. In this brief note, however, we suggest (based on a learning curve experiment) that finding a Preference Order for a new domain may not be so difficult after all, as long as one has access to a handful of human-produced descriptions collected in a semantically transparent way. We argue that this is due to the fact that it is both more important and less difficult to get a good ordering of the head than of the tail of a Preference Order.

Keywords: Generation/production of referring expressions; Evaluation metrics for generation algorithms; Incremental algorithm; Psycholinguistics; Reference; Learning curve experiments

1. Introduction

When speakers use a description to refer to an object (‘‘the green sofa’’) or a person (‘‘the tall man’’), they have to determine which properties to include so that an addressee knows which object or person is intended. Clearly, there are many ways in which a sofa can be distinguished from other furniture items (‘‘the large sofa,’’ ‘‘the sofa left of the modern-looking chair,’’ etc.), and there are even more possibilities when we consider references to persons. Various algorithms for the Generation of Referring Expressions (GRE) have been proposed that compute which set of properties distinguish a target, where properties

Correspondence should be sent to Emiel Krahmer, Tilburg Center for Cognition and Communication (TiCC), School of Humanities, Tilburg University, P.O. Box 90153, NL-5000 LE, Tilburg, The Netherlands. E-mail: e.j.krahmer@uvt.nl.

themselves are often represented as attribute-value pairs, such as ⟨ COLOR, *green* ⟩, indicating that the target has the value *green* for the attribute COLOR. In a seminal study, Dale and Reiter (1995) present and compare various algorithms that accomplish this task. One algorithm, which they call Greedy Heuristic (GR), relies on discriminatory power: It chooses properties one by one, at each iteration selecting that property of the intended referent that excludes most of the distractors not previously ruled out. An alternative that Dale and Reiter discuss is the Incremental Algorithm (IA), which relies on the assumption that some attributes (e.g., COLOR) are more preferred than others (e.g., SIZE). This notion is formalized using a preference order (PO), which is a list of attributes through which the IA iterates, selecting an attribute-value pair if it helps distinguishing the target from one or more of the distractors. Dale and Reiter point out that POs may vary from one domain to another, and should be determined empirically.

Recently, van Deemter et al. (2012) presented an extensive evaluation of the various algorithms discussed in Dale and Reiter (1995). In their paper, van Deemter and colleagues emphasize the crucial role of the PO when evaluating the IA. They show that, both in a furniture and in a people domain, the IA outperforms the GR algorithm with a ''good'' PO, while the opposite holds when the IA relies on a ''bad'' PO. This outcome naturally raises the question ''how 'good' POs might be selected'' (van Deemter et al., 2012, p. 6). The authors point out that systematically testing all POS for a given domain quickly becomes impractical, since for a domain where objects can be described using n different attributes, there are $n!$ POs to consider. Based on considerations such as these, they conclude that ''someone who is looking for a GRE algorithm for a previously unstudied application domain might do better choosing GR (…), instead of an unproven version of the IA'' (van Deemter et al., 2012, p. 31).

However, it seems to us that it should first be determined how difficult it really is to find a ''good'' PO. Perhaps only a handful or even only one human-produced description is needed to make an educated guess about the PO for a certain domain. One way to find out is by running a learning curve experiment, in which randomly selected sets of human-produced referring expressions (increasing in size) are used to determine the ranking of attributes on a PO, after which the IA (with the various, resulting POS) can be evaluated in the same way as was done by van Deemter et al. (2012). In this letter, we report on exactly such an experiment, thereby extending the findings of van Deemter and colleagues.

2. Experiment

2.1. Method

For our experiment we relied on the TUNA data-set, extensively described in van Deemter et al. (2012). This semantically transparent corpus contains referring expressions for two domains (furniture and people), and it has been used for comparative evaluation in the REG Challenges (2007–2009). To determine POs, we used 165 furniture descriptions and 136 people descriptions from the training set of the REG Challenge 2009. For evaluating the IA, we

used 38 furniture descriptions and 38 people descriptions from the 2009 REG Challenge development set.

We used randomly selected subsets for each domain of the training set, with set sizes of 1, 5, 10, 20, 30, 40, and 50 descriptions, as well as the entire set. Because the accidental composition of a training set may strongly influence the results, especially for small training sets, we created five different sets of each size. The training sets were built up in a cumulative fashion: We started with five randomly selected sets of size 1, then added four randomly selected descriptions to each of them to create five sets of size 5, etc.

The rank of an attribute on a PO was determined by counting the number of descriptions that include this attribute in a given set. Van Deemter et al. (2012) likewise relied on frequencies (as well as on psycholinguistic evidence) to determine promising POs (which, incidentally, turned out to be easier to find for the furniture than the people domain), but, crucially, where they only computed frequencies over their entire development set, we specifically consider subsets (of increasing sizes).

Consider determining a PO based on a set of one description, say ''the large chair.'' Two attributes are mentioned in this description (SIZE and TYPE) and hence have frequency 1; these are placed at the head of the PO. Other attributes (such as COLOR and ORIENTATION) are not mentioned (frequency 0), and these are placed at the tail of the PO. This gives a partial ordering of attributes. However, the IA requires a complete ordering, so when two or more attributes occur in a single training set with the same frequency, we decided to rank them alphabetically (clearly, as set size increases, frequencies will more often be enough to completely rank all attributes). The resulting PO for our example would be ⟨ SIZE, TYPE, COLOR, ORIENTATION ⟩. Note that like Dale and Reiter (1995), but contrary to van Deemter et al. (2012) we check whether TYPE is in the description after running the IA and include it when it was not added earlier. Nothing hinges on this for the experiment we describe here, however.

Once a PO has been determined, we run the IA with this PO on the test set and evaluate the performance in terms of Dice and PRP, as van Deemter et al. (2012) do. Dice measures set overlap between a generated set of properties and those produced by a human speaker, ranging from 0 (no overlap) to 1 (complete overlap); PRP is the perfect recall percentage, the proportion of times when the IA achieves a Dice score of 1.

2.2. Results

Table 1 summarizes the results and reveals a clear picture. For the Furniture domain, a set size of five descriptions is sufficient to reach top performance both in terms of Dice and PRP. For the more complex People domain the results are slightly more varied, and increasing set sizes generally give numerically higher scores. However, it is important to note that even here the first few descriptions clearly have the biggest impact; moving from set size 1 to set size 10, for example, yields a PRP improvement of 0.137, while moving from set size 10 to the *entire* set gives a PRP improvement of 0.105. For Dice, this is even clearer: From set size 1 to 10 gives an increase of 0.16, while moving from 10 to the entire set improves the Dice score with only 0.04.

Table 1
Dice and PRP scores for the furniture and people domains as a function of training set size; the entire set consists of, respectively, 165 and 136 descriptions

Size	Furniture		People	
	Dice	PRP	Dice	PRP
01	0.764	0.368	0.519	0.074
05	0.829	0.553	0.605	0.158
10	0.829	0.553	0.682	0.211
20	0.829	0.553	0.710	0.221
30	0.829	0.553	0.682	0.153
40	0.829	0.553	0.716	0.263
50	0.829	0.553	0.718	0.279
all	0.829	0.553	0.724	0.316

To test for statistical significance, we ran separate analyses of variance (one for each domain) with set-size as a repeated measure, and Dice and PRP as the dependent variables. Planned post hoc pairwise comparisons were made to test which set-size is the smallest that does not perform significantly different from the entire set (we call this the ''ceiling''); we report results both for the standard Bonferroni method, which corrects for multiple comparisons, and for the less strict LSD method from Fisher, which does not. Note that with the more strict Bonferroni method we are inherently less likely to find statistically significant differences between set sizes, meaning that we can expect to reach a ceiling earlier than with the LSD method.

We found statistically significant effects of set size for both the Furniture (Dice: $F(7, 259) = 13.81, p < .001, \eta_p^2 = 0.27$ and PRP: $F(7, 259) = 34.31, p < .001, \eta_p^2 = 0.48$) and the People domains (Dice: $F(7, 259) = 18.47, p < .001, \eta_p^2 = 0.33$ and PRP: $F(7, 259) = 5.99, p < .001, \eta_p^2 = 0.14$), indicating that the number of descriptions from which a PO is derived influences the performance of the IA both in terms of Dice and of PRP. Crucially, the post hoc tests confirm that, generally speaking, small set sizes are sufficient to reach ceiling performance. For the furniture domain, we (clearly) reached the ceiling at set-size 5. For the People domain, the ceiling for Dice is reached at set-size 10 (both with the Bonferroni and the LSD method), while for PRP the ceiling is reached at set-size 40 using the LSD method, and 1 with the Bonferroni method.

3. Concluding remarks

Our experiment shows that even based on a small set of human-produced descriptions we can make an informed guess about what a ''good'' PO is. With relatively small sets of descriptions (much smaller than the entire TUNA corpus) we obtain results that are not significantly different from the best performing variant of the IA. Arguably, these results allow two important, related conclusions: (a) Even though, in theory, there are indeed $n!$

different possible POS of *n* attributes, it is both more important and less difficult to get a ''good'' ordering for the head of the PO than for its tail. And (b) relatively many human-produced descriptions will include attributes that are preferred, so that it does not really matter how dispreferred attributes are ordered. It is worth pointing out that our approach still requires a semantically balanced corpus. However, the evaluation results suggest that a surprisingly small one may be sufficient. Of course, it should be kept in mind that these conclusions, while intuitive, are based on data from the relatively simple domains studied by van Deemter et al. (2012). Moreover, here we have only considered references to single targets, while van Deemter et al. (2012) also consider references to sets, for which determining a PO might well be more complicated.

It will certainly be interesting to see what happens when the field of referring expression generation moves toward more complex references and more open-ended domains. For now, we conclude that someone who is looking for a GRE algorithm for a new application domain might consider collecting a handful of human-produced descriptions, before discarding the IA in favor of another algorithm.

Acknowledgments

Krahmer and Koolen received financial support from the Netherlands Organization for Scientific Research (Vici grant 27770007). Thanks to Albert Gatt for allowing us to use his implementation of the Incremental Algorithm and his evaluation scripts, and to Kees van Deemter and Albert Gatt for useful discussions and comments on an earlier version of this text.

References

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

van Deemter, K, Gatt, A, van der Sluis, I., & Power, R. (2012). Generation of referring expressions: Assessing the incremental algorithm. *Cognitive Science*, doi: 10.1111/j.1551-6709.2011.01205.x