Salience in the Generation of Multimodal Referring Acts

Paul Piwek
Centre for Research in Computing, The Open University
Milton Keynes, United Kingdom
p.piwek@open.ac.uk

ABSTRACT

Pointing combined with verbal referring is one of the most paradigmatic human multimodal behaviours. The aim of this paper is foundational: to uncover the central notions that are required for a computational model of multimodal referring acts that include a pointing gesture. The paper draws on existing work on the generation of referring expressions and shows that in order to extend that work with pointing, the notion of salience needs to play a pivotal role. The paper starts by investigating the role of salience in the generation of referring expressions and introduces a distinction between two opposing approaches: salience-first and salience-last accounts. The paper then argues that these differ not only in computational efficiency, as has been pointed out previously, but also lead to incompatible empirical predictions. The second half of the paper shows how a salience-first account nicely meshes with a range of existing empirical findings on multimodal reference. A novel account of the circumstances under which speakers choose to point is proposed that directly links salience with pointing. Finally, this account is placed within a multi-dimensional model of salience for multimodal reference.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Language generation*; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

Keywords

Referring Expressions, Pointing Gestures, Salience, Incremental Algorithm

1. INTRODUCTION

Researchers on human pointing gestures have for a long time observed that pointing is essentially a means to "reorient the attention of another person so that an object becomes the shared focus of attention" (Butterworth [2]). In other words, pointing can be viewed as one way to change the salience of an object.[1] Somewhat surprisingly, this insight does not seem to have had any counterpart in computational models for generating multimodal referring expressions. These accounts do not feature the notion of salience and typically treat pointing as a fallback strategy. For example, Lester et al. [11] proposes to only use a pointing act, if a pronoun doesn't suffice to identify the target, and Claassen [4] introduces an algorithm which only uses pointing if no purely verbal means of identification is possible. Similarly, Van der Sluis and Krahmer [16] describe an algorithm that only uses a pointing act if a purely verbal referring act becomes too complex. More recently, Krahmer and Van der Sluis [10] describe an algorithm for multimodal reference that does not view pointing as a fallback strategy. Their algorithm assigns costs to the properties that are included in a referring expression. Pointing is modelled as just another property: a pointing act identifies a subset of objects in the domain. A graph-based algorithm is employed to find the cheapest combination of properties for referring to an object. But again, salience plays no role in this account. Also accounts of the interpretation of multimodal referring acts have typically not related salience with pointing; e.g., Choumane and Siroux [3], who do model visual salience, view pointing acts strictly as designating an object.

The aim of this paper is to unpick the relation between salience and pointing and lay the foundations for a computational account. The starting point is another look at the role of salience in the generation of referring expressions. I will distinguish between two opposing approaches for dealing with salience: salience-first and salience-last accounts, and argue that these differ not only in computational efficiency, as has been pointed out previously, but also lead to different empirical predictions. The second half of the paper shows how a salience-first account nicely meshes with a range of existing empirical findings on multimodal reference. I propose

[1]Salience is used here in a very broad sense, following, for example, Theune [14] and Krahmer and Theune [9]. An object can be salient because it has been pointed at, referred to, or also, for instance, because of its colour (which may be different from the objects around it) or its size. Some may prefer to replace the term 'salience' with 'accessibility', e.g., as defined on Page 699 of Kahneman [7]: "[...] accessibility—the ease (or effort) with which particular mental contents come to mind. The accessibility of a thought is determined jointly by the characteristics of the cognitive mechanisms that produce it and by the characteristics of the stimuli and events that evoke it. [...] the determinants of accessibility subsume the notions of stimulus salience, selective attention, specific training, associative activation, and priming."

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

a novel account of the circumstances under which speakers point that directly links salience with pointing. Finally, I introduce a multi-dimensional model of salience.

2. SALIENCE: FIRST OR LAST?

Throughout this paper, Dale and Reiter's [5] incremental algorithm (IA) is used as a starting point. The IA works on the assumption that there is a universe or domain of objects $\mathcal{U}$ which includes a target r, the object the speaker intends to refer to. In order to refer to r, the speaker constructs a description D which expresses a set of properties $P_1, \ldots, P_n$ such that the intersection of these properties equals $\{r\}$. In other words, the description is such that it uniquely identifies r. Each property is treated extensionally[2] as a subset of $\mathcal{U}$ and properties are organized as belonging to attributes (e.g., the properties *red*, *green*, ... are associated with the attribute *colour*). Attributes are ordered, where the ordering indicates which attributes are preferred for constructing a description.

The algorithm works as follows: it starts with the empty description $D = \emptyset$ and a context set C which is initialized with the domain: $C = \mathcal{U}$, and iterates through the ordered list of attributes. The algorithm fails if the end of the list is reached. On each iteration, the following steps are taken:

1. The best property P belonging to the current attribute is selected, i.e., the property P which has the smallest non-empty intersection with C and includes r.
2. **If** $C - P \neq \emptyset$ (where $C - P$ stands for the set of objects in C that are ruled out by P), **then**: $C = C \cap P$ and $D = D \cup \{P\}$
3. **If** $C = \{r\}$ **then**: return D, unless D includes no property from the top-ranked attribute, in which case add an appropriate property from this attribute to D and return the result.[3]

There are two principal ways to add salience to this account. They can be compared most easily by assuming that salience S_r is a property, i.e., a subset of $\mathcal{U}$ that can be computed if we know the salience value of each of the objects in $\mathcal{U}$ and the identity of the target r:

> S_r, the *salience property for* r, is the set of objects whose salience value is above some threshold value which is defined as the salience value of r minus a confidence interval (see Figure 1).

Note that at this point we remain agnostic about how individual salience values are computed. We address this issue in Section 4.

In *salience-first* accounts, IA is started by initializing C with $S_r (\subseteq \mathcal{U})$ instead of $\mathcal{U}$: the idea is to find a description which distinguishes r from the objects in $\mathcal{U}$ that, given a confidence interval, are at least as salient as r itself. Alternatively, *salience-last* accounts modify iteration step 3:

[2]In order to avoid notational clutter, we use P to refer both to the name of a property and the property itself, rather than writing $\|P\|$ for the property.

[3]Thus, for example, in a domain consisting only of triangles, the algorithm will produce the description 'the blue triangle' to identify a blue triangle, even though 'triangle' is strictly speaking not required to identify the target.

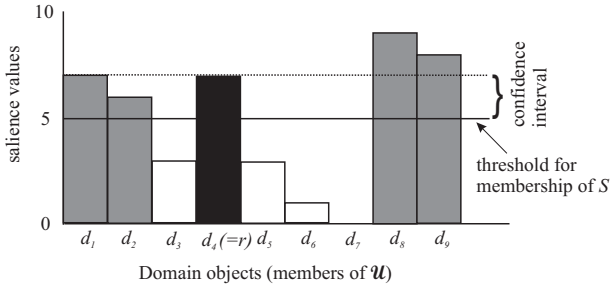

Figure 1: A bar chart depicting for each object in some domain $\mathcal{U}$ the corresponding salience value. The target is represented by a black bar and the other members of the salience property S_r are distinguished by their grey colour.

the condition $C = \{r\}$ is replaced by $C \cap S_r = \{r\}$. Thus, at the end of each iteration it is checked whether r is the most salient object which fits the description D. Whereas, for example, Theune [14] and Van Deemter and Krahmer [15] propose essentially salience-first accounts, Kelleher et al. [8] and Krahmer and Theune [9] describe salience-last algorithms. The former point out that their approaches are to be preferred on computational grounds; by removing from $\mathcal{U}$ all objects that are not a member of S_r, the algorithm, at each step, has to inspect a smaller C than in any salience-last approach. A further possible reason for preferring salience-first is its cognitive plausibility (Van Deemter and Krahmer mention its 'naturalness', though they do not expand on this). Here, we draw attention to a novel observation: salience-first and salience-last accounts lead to different empirical predictions.

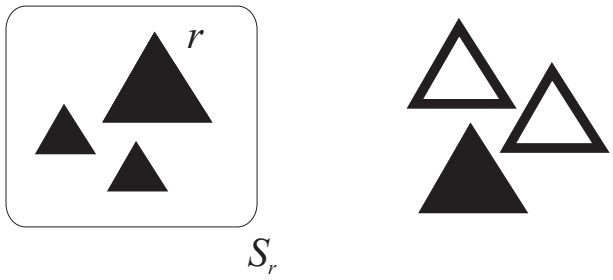

Figure 2: A domain with several triangles. The set of triangles enclosed by the box is the salience property S_r for r

Consider Figure 2 and let us assume that the attributes are ordered as follows: *shape*, *colour*, *size*.[4] The salience-first approach results in $D = \{big, triangle\}$: C is restricted to the set of salient objects (the ones within the enclosed area). Since all objects are triangles, on the first iteration no property is added to D. On the second iteration, no property is added either (since all salient objects have the same colour). On the third and final iteration, the property

[4]For this particular example we need the ordering that we provided, but it is straightforward to create examples of the same type based on different orderings.

big is added which distinguishes r from the other objects in C. Finally, $D \cup \{triangle\}$ is returned, which can be realized as for example 'the big triangle'. Salience-last, in contrast, results in $D = \{black, big, triangle\}$. This is a consequence of the fact that in the second iteration, the test on whether to include *black* is: **a**) Does it include r? *Yes.* **b**) Does it rule out any objects from $\mathcal{U}$ (rather than $S_r (\subseteq \mathcal{U})$)? *Yes, the two white triangles.*

3. WHEN TO POINT?

In contrast with the accounts of pointing discussed in Section 1, here we put forward a model for multimodal reference which establishes a direct link between pointing and salience, and more specifically salience-first accounts. The basic ingredients of this approach are:

1. Pointing is a way of making the set of objects that have been pointed at maximally salient.
2. Assuming that the target r is a member of the set of objects that the speaker pointed at, the pointing act causes S_r to be identical with the set of objects that the speaker pointed at.
3. In accordance with the salience-first version of the Incremental Algorithm, S_r (the salience property for r) is used to initialise the context set C, and a description is generated relative to this set.

This tells us what the effect of pointing is. We propose that the decision *when* to point is captured by the following rule:

> SALIENCE-BASED POINTING RULE: If, as a result of pointing, the size of the context set $C = S_r$ for target r can be reduced, then point.[5]

This account is grounded in the following empirical findings:

1. The decision whether to point is correlated with the salience of the target: pointing is preferred when the target is *not* salient, i.e., when S_r is big relative to the domain $\mathcal{U}$ (Piwek [13]).
2. When the target is pointed at, on average the number of properties used in the description is smaller (Piwek [13]).
3. Levelt et al. [12] and De Ruiter [6] found that the onset of pointing gestures *precedes* their spoken affiliates. This is compatible with the model proposed here, where a speaker *first* decides whether to point and then constructs a verbal description.

Let us compare this approach with the one based on costs advocated by Krahmer and Van der Sluis [10] (see Section 1). Consider Figure 3. Using the cost assignments provided in [10], we can calculate that the optimal description of the target r is 'the small black triangle' (cost 2.25). This description is cheaper than 'this triangle' + pointing (cost 3).

[5]This rule may need to be refined for situations where the size of S_r is very small to start with: we may need to add a condition to the rule requiring that $S_r > c$, where c is a constant that will need to be determined empirically. Also, the degree to which S_r is reduced may need to be taken into account.

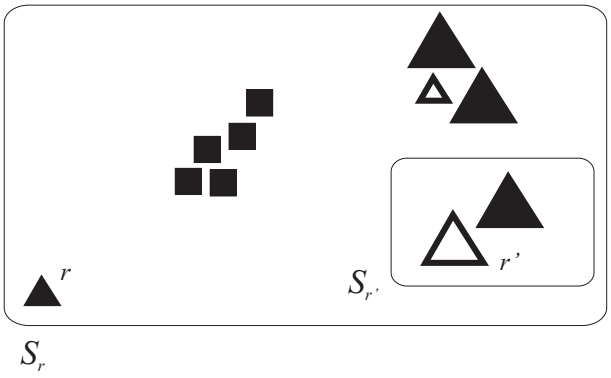

Figure 3: Example of a domain; two targets, r and r', are marked together with their respective salience properties, S_r and $S_{r'}$

Of course, with a different cost assignment (e.g., making verbal properties more expensive and pointing cheaper) the solution changes. More importantly, however, what the cost model does not capture is that pointing is a fast way to reduce S_r. Compare this with a reference to the target r'. Here we have a small $S_{r'}$ to start with, and pointing may not help from where the speaker is standing: assuming the speaker remains stationary, s/he may only be able to point at a set of objects that is equal to or bigger than $S_{r'}$. The cost-based model ignores these considerations.

4. DIMENSIONS OF SALIENCE

So far we have not dealt with the detail of how to compute the salience values that determine S_r. We have suggested that pointing can change salience values. Also, there is ample literature on how verbal reference affects salience. Usually the idea is that the more recent an object was referred to, the more salient it is. In a visually shared domain, spatial relations between objects can also influence salience. In particular, an object that is salient directs attention to itself and the spatial region around it. Consequently, the salience of the objects in its vicinity get a boost - here we will call this implied spatial salience. Beun and Cremers [1] have found that speakers exploit spatially implied salience in that they usually produce (first-mention) descriptions that only distinguish the target from the most salient object and objects that are spatially implied by (i.e., close to) it.

For each of the aforementioned types of salience, we propose to introduce a separate dimension modelled as a function: p (pointing dimension), v (verbal reference dimension) and i (implied spatial dimension). Each function, when applied to a specific object x returns an integer from $[0 - 10]$. We also define the aggregate salience value of an object as: $s(x) = max(p(x), v(x), i(x))$. The dynamics of p, i and v are given by the following equations which relate the dimensions to states (indicated by subscripts):

- $p_0(x) = v_0(x) = i_0(x) = 0$
- $p_S(x) = \begin{cases} 10 \text{ if } x \text{ is pointed at between } S-1 \text{ and } S \\ \text{else } 0 \end{cases}$

- $v_S(x) = \begin{cases} 10 \text{ if condition } \dagger(x) \text{ holds.} \\ v_{S-1}(x) - 1 \text{ if not } \dagger(x) \text{ and} \\ \quad v_{S-1} > 0 \ \& \ \neg\exists y : p_{S-1}(y) = 10 \\ v_{S-1}(x) \text{ if not } \dagger(x) \ \& \ \exists y : p_{S-1}(y) = 10 \\ \text{else } 0 \end{cases}$

- $i_S(x) = \begin{cases} 8 \text{ if } \exists y : v_S(y) = 10 \text{ and} \\ \quad x \text{ spatially implies } y \\ \text{else } 0 \end{cases}$

Here, $\dagger(x) \Leftrightarrow x$ is referred to between $S-1$ and S.
The equations can be seen at work in Figure 4. This figure depicts a sequence of states for a universe of two objects, d_1 and d_2. Although in this model states are temporally ordered, transitions between states can take place in parallel, as long as a transition to a later state is never completed before the transitions to the states preceeding.

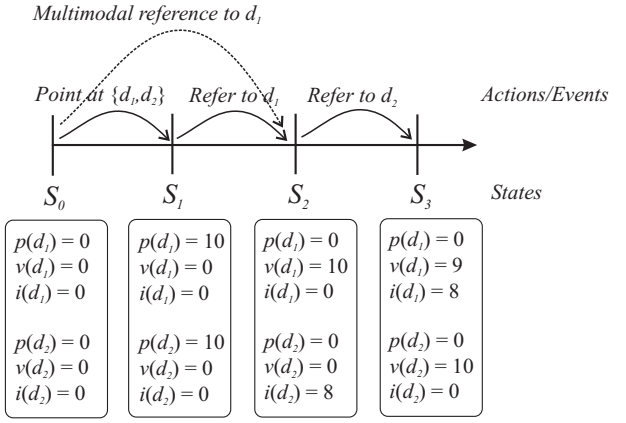

Figure 4: Example of how salience values change as a result of pointing and reference. v, i and r stand for the three dimensions of salience: the pointing, implied spatial and verbal reference salience dimension.

5. CONCLUSIONS

We have proposed a novel account of when to include pointing in a referring act. The proposal follows the insight from the study of human pointing gestures that pointing is primarily a means for changing the salience of objects. Our account is framed in terms of a salience-first algorithm. We demonstrated that salience-first differs not only in computational efficiency but also in empirical predictions from salience-last approaches. The proposal is grounded in a number of empirical findings about human multimodal referring acts and will hopefully provide a fruitful starting point for the development of multimodal interactive systems.

6. REFERENCES

[1] R.J. Beun and A. Cremers. Object reference in a shared domain of conversation. *Pragmatics & Cognition*, 6(1/2):121–152, 1998.

[2] G. Butterworth. Pointing is the royal road to language for babies. In S. Kita, editor, *Pointing: Where Language, Culture and Cognition Meet*, pages 9–34. Lawrence Erlbaum Associates, Mahwah, NJ, 2003.

[3] Ali Choumane and Jacques Siroux. Knowledge and Data Flow Architecture for Reference Processing in Multimodal Dialogue Systems. In *2008 Conference on Multimodal Interfaces (ICMI'08)*, Crete, Greece, 2008. ACM.

[4] W. Claassen. Generating referring expressions in a multimodal environment. In R. Dale et al., editor, *Aspects of Automated Natural Language Generation*. Springer Verlag, Berlin, 1992.

[5] R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263, 1995.

[6] J.P. de Ruiter. *Gesture and Speech Production*. PhD thesis, Max Planck Institute, Nijmegen, 1998.

[7] D. Kahneman. A perspective on judgement and choice: Mapping bounded rationality. *American Psychologist*, 58(9):697–720, 2003.

[8] J. Kelleher, F. Costello, and J. van Genabith. Dynamically structuring, updating and interrelating representations of visual and linguistic discourse context. *Artificial Intelligence*, 167:62–102, 2005.

[9] E. Krahmer and M. Theune. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing*, pages 223–264, Stanford University, 2002. CSLI.

[10] E. Krahmer and I. van der Sluis. A new model for the generation of multimodal referring expressions. In *Proceedings European Workshop on Natural Language Generation (ENLG2003)*, Budapest, Hungary, 2003.

[11] J. Lester, J. Voerman, S. Towns, and C. Callaway. Deictic Believability: Coordinating gesture, locomotion, and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13(4-5):383Ű414, 1999.

[12] W.J. Levelt, G. Richardson, and L. Heij. Pointing and voicing in deictic expressions. *Journal of Memory and Language*, 24:133–164, 1995.

[13] P. Piwek. Modality choice for generation of referring acts: Pointing versus describing. In *Proceedings of Workshop on Multimodal Output Generation (MOG 2007)*, pages 129–139, Aberdeen, Scotland, January 2007.

[14] M. Theune. *From Data to Speech: Language Generation in Context*. PhD thesis, Eindhoven University of Technology, 2000.

[15] K. van Deemter and E. Krahmer. Graphs and Booleans: on the generation of referring expressions. In H. Bunt and R. Muskens, editors, *Computing Meaning*, volume 3. Kluwer, Dordrecht, 2006.

[16] I. van der Sluis and E. Krahmer. Generating Referring Expressions in a Multimodal Context: An empirically motivated approach. In *Selected Papers from the 11th CLIN Meeting*. Rodopi, Amsterdam, 2001.