Serial Dependency: Is It a Characteristic of Human Referring Expression Generation?

Jette Viethen[1,2]
jette.viethen@mq.edu.au

Robert Dale[2]
robert.dale@mq.edu.au

Markus Guhe[3]
m.guhe@ed.ac.uk

[1]TiCC
University of Tilburg
Tilburg, The Netherlands

[2]Centre for Language Technology
Macquarie University
Sydney, Australia

[3]School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

A key characteristic of many existing referring expression generation (REG) algorithms is **serial dependency**: attributes are selected for inclusion one at a time, and the decision to include each attribute is dependent on the discriminatory ability of the set of attributes that have already been selected so far. We use a machine learning approach to explore whether serial dependency is a characteristic of human referring expression generation. Our results show that models in which attributes are chosen in a serially dependent fashion does not perform better than one where their inclusion depends only on other factors. The results also suggest that the visual salience of an attribute might be more important than its discriminatory power.

Introduction

Most work on referring expression generation is focussed on content selection: the choice of the properties of the intended referent that should be included in the referring expression (see, for example, Dale and Reiter, 1995; Krahmer et al., 2003; van Deemter, 2006). Many existing referring expression generation (REG) algorithms follow the principle of **serial dependency**: attributes are selected for inclusion one at a time, and the decision to include each attribute is dependent on the discriminatory ability of the set of attributes that have already been selected so far. This observation applies in particular to Dale's (1989) Greedy Algorithm, Dale and Reiter's (1995) Incremental Algorithm (IA), and a wide range of other reported algorithms which are based on these, such as those described in (Gardent, 2002) and (Krahmer and Theune, 2002).

In (Dale and Viethen, 2009) we suggested that the adherence to serial dependency mght be one reason why existing algorithms are not able to generate the full range of different referring expressions that humans produce. As an alternative, we proposed that attributes could be considered individually and in parallel, making their inclusion in a referring expression dependent only on factors external to the referring expression currently under construction.

In this paper we explore the question of whether serial dependency is a property of human referring expression generation. We compare the output of a number of machine-learned models for REG to referring expressions produced by human speakers with the aim of determining which of these models is best able to replicate the human data. The models are derived by training decision trees which make binary decisions as to whether or not a specific attribute of a target referent should be included in a referring expression.

Our models use two types of features: **chaining features** provide information about the discriminatory power of both the referring expression constructed so far and the attribute currently under consideration. Discriminatory power is determined by the number of distractor entities that can be ruled out by using an attribute or a set of attributes. These features capture the information required for serial dependency. **Non-chaining features** represent aspects of the broader discourse and visual context surrounding the target referent, and are independent of the ongoing process of the production of a referring expression.

We hypothesise that, if serial dependency plays a role in the generation of referring expressions, then the models that use chaining features should achieve a closer match to the human-produced data than the models that only use non-chaining features. Our experiments demonstrate that the chaining features do not contribute significantly to an accurate model of human production of referring expressions, lending support to the view that the property of serial dependency that is central to the IA and similar algorithms does not accurately reflect the way in which humans generate referring expressions.

In the following, we first introduce the data set we use for our experiments, and then describe the machine learning features and resulting REG models in some detail. Finally, we present the test results for the different models and draw some conclusions.

The Data

The iMAP Corpus Louwerse et al. (2007) is a collection of 256 dialogues between 32 participant-pairs who contributed 8 dialogues each. Both participants had a map of the same environment, but one participant's map showed a route winding its way between the landmarks on the map; see Figure 1. The task was for this participant (the instruction giver, IG) to describe this route in such a way that their partner (the instruc-

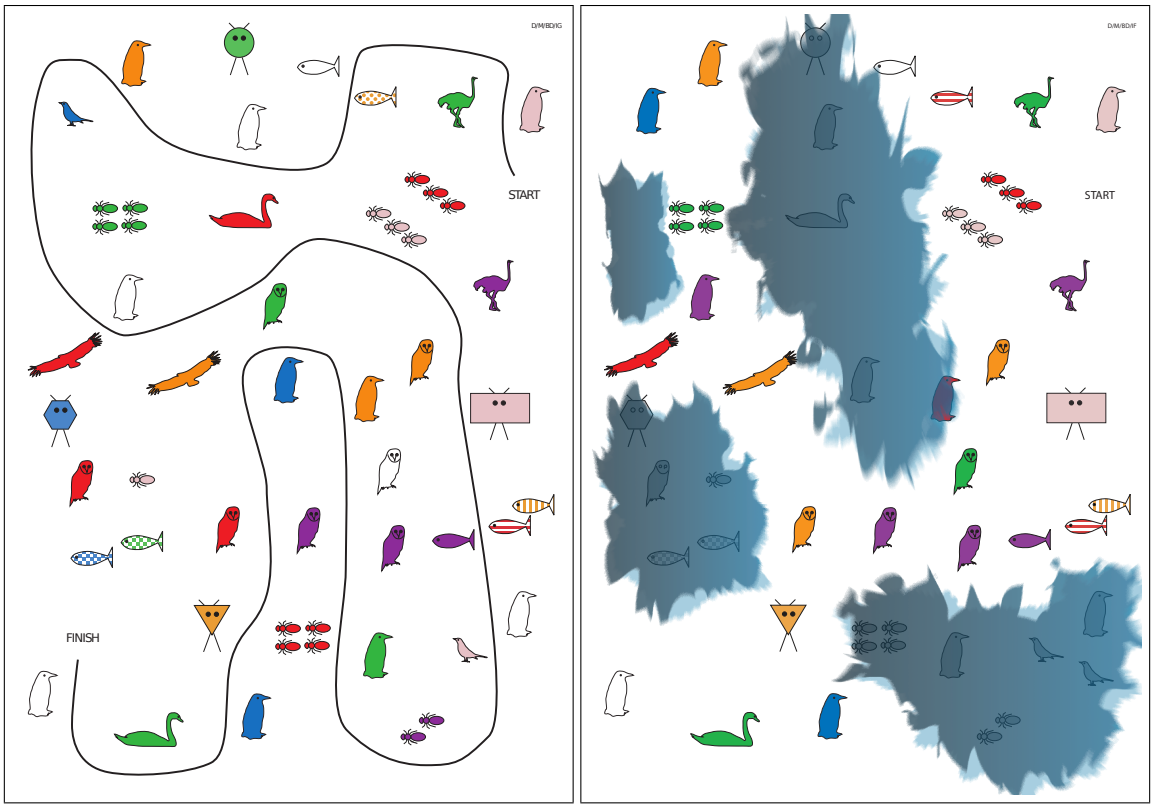

Figure 1: An example pair of maps.

tion follower, IF) could draw it onto their map; this was complicated by some discrepancies between the two maps, such as missing landmarks, the unavailability of colour in some regions due to ink stains, and small differences between some landmarks.

The landmarks differ from each other in type, colour, and one other attribute, which is different for each type of landmark. For example, there are different *kinds* of birds (eagle, ostrich, penguin, ...); fish differ by their *patterns* (dotted, checkered, plain, ...), aliens have different *shapes* (circular, hexagonal, ...), and bugs appear in small clusters of differing *numbers*. In addition to these inherent attributes of the landmarks, participants used spatial relations to other items on the map.

From the corpus as a whole we extracted 34,127 referring expressions. We removed from the data any annotation that was not concerned with the four landmark attributes, type, colour, relation, or the landmark's other distinguishing attribute; so, for example, we removed 'semantically empty' head nouns such as *thing* or *set*. Ordinal numbers that were annotated (in our view, incorrectly) as the use of the number attribute were re-tagged as spatial relations, as these usually described the position of the target within a line of landmarks.

As a result of the removal of annotations not pertaining to the use of the four landmark attributes, 2,785 referring expressions had no annotation left; we removed these instances from the final data set. We also do not attempt to replicate the remaining 5,552 plural referring expressions or the 3,062 pronouns found in the corpus.[1] Similarly, we excluded 2,586 referring expressions that made use of a spatial relation, because it would be difficult to determine the discriminatory power of a relation in this setting. However, we do include all of these instances in the feature extraction step, on the assumption that they might impact on the content of subsequent references. This resulted in a final data set of 20,141 referring expressions, made up of 5,936 initial references and 14,205 subsequent references.

We capture the presence or absence of the three attributes, type, colour and other, in a referring expression by means of a **content pattern**, abstracting away from lexico-syntactic characteristics of the referring expression as well as the ac-

[1]The additional issues that arise in generating plural references and deciding when to use pronouns considerably complicate the problem; see Gatt (2007); Horacek (2004).

Content Pattern	Count	Proportion
⟨other⟩	7561	37.5%
⟨other, type⟩	5975	29.7%
⟨other, colour⟩	2364	11.7%
⟨other, colour, type⟩	1954	9.7%
⟨colour⟩	1029	5.1%
⟨type⟩	662	3.3%
⟨colour, type⟩	596	3.0%
Total	20,141	

Table 1: The seven content patterns by frequency.

tual value of each attribute. So, for example, if a particular reference appears as the noun phrase *the blue penguin*, annotated semantically as ⟨blue, penguin⟩, then the corresponding content pattern is ⟨colour, kind⟩. Our aim is to replicate the content pattern of each referring expression found in the corpus. Table 1 lists the seven content patterns that occur in our data set in order of frequency.

The Models

For each of the three attributes (type, colour and other) we trained a decision tree that makes a binary decision as to whether or not to include that attribute. The output of the three trees is then combined into a content pattern, which can be compared to the content pattern of the corresponding human-produced description in our corpus.

For the unchained model, the trees are only given access to a large set of **non-chaining** features. This set consists of three subsets: TradREG features, which are based on the concerns that traditional REG work is focussed on (Table 2); Alignment features, which are based on work in psycholinguistics that suggests the people often re-use the form and content of previous referring expressions (see for example, Clark and Wilkes-Gibbs, 1986; Carroll, 1980; Brennan and Clark, 1996; Pickering and Garrod, 2004; Goudbeek and Krahmer, 2010; Table 3); and Ind features, which are a collection of theory-independent features (Table 4).[2] In (Viethen et al., 2011) we found that, at least for subsequent reference, using the complete set of these features outperforms models based on any of the subsets. We therefore use the full feature set in the experiments presented here. Jordan and Walker (2005) had a similar finding on a different data set.

To derive chained models, we trained a 'chain' of attribute-specific decision trees for each of the six possible sequences of the three attributes. In this case, each decision tree was given access to chaining features which provide information about the referring expression constructed so far. Of central importance for these features is the concept of **discriminatory power**, which is defined as the ratio between the number of distractor entities to which a given attribute or set of attributes does not apply and the total number of distractors. Here are the three types of chaining features that we use:

[2] In these tables, *Att* is an abbreviatory variable that is instantiated once for each of the three attributes type, colour, and the other distinguishing attribute of the landmark. The abbreviation LM stands for landmark.

TradREG Features (Visual)	
Count_Vis_Distractors	number of visual distractors
Prop_Vis_Same_*[Att]*	proportion of visual distractors with same *Att*
Dist_Closest	distance to the closest visual distractor
Closest_Same_*[Att]*	has the closest distractor the same *Att*?
Dist_Closest_Same_*[Att]*	distance to the closest distractor of same *Att* as target
Cl_Same_type_Same_*[Att]*	has the closest distractor of the same type also the same *Att*?
TradREG Features (Discourse)	
Count_Intervening_LMs	number of other LMs mentioned since the last mention of the target
Prop_Intervening_*[Att]*	proportion of intervening LMs for which *Att* was used, and which have the same *Att* as target

Table 2: The TradREG feature set.

1. **DP_*Att*** represents the discriminatory power of attribute *Att* at the time at which *Att* is considered for inclusion. We use two versions of this feature, one pertaining to visual distractors (the surrounding landmarks on the map) and one to discourse distractors (landmarks that have recently been mentioned in the dialogue). Most existing REG algorithms, including the IA and the Greedy Algorithm, make their decision as to whether to include an attribute *Att* dependent on its discriminatory power.

   The DP_*Att* features are related to the TradREG features Prop_Vis_Same_*Att* and Prop_Intervening_*Att*, but with the important difference that their values are computed at run time, taking into account that the size of the total distractor set might already be reduced by attributes that have already been included. This makes the DP_*Att* features chained versions of the two TradREG features.

2. **DP_RE** records the discriminatory power of the referring expression built so far. Existing algorithms stop adding attributes as soon as the discriminatory power of the referring expression reaches 1, which means that all distractors are ruled out. Again, we use both visual and discourse variants

Alignment Features (Recency)	
Last_Men_Speaker_Same	who made the last mention of target?
Last_Mention_*[Att]*	was *Att* used in the last mention of target?
Dist_Last_Mention_Utts	distance to the last mention of target in utterances
Dist_Last_Mention_REs	distance to the last mention of target in REs
Dist_Last_*[Att]*_LM_Utts	distance in utterances to last use of *Att* for target
Dist_Last_*[Att]*_LM_REs	distance in REs to last use of *Att* for target
Dist_Last_*[Att]*_Dial_Utts	distance in utterances to last use of *Att*
Dist_Last_*[Att]*_Dial_REs	distance in REs to last use of *Att*
Dist_Last_RE_Utts	distance to last RE in utterances
Last_RE_*[Att]*	was *Att* mentioned in the last RE?
Alignment Features (Frequency)	
Count_*[Att]*_Dial	how often has *Att* been used in the dialogue?
Count_*[Att]*_LM	how often has *Att* been used for target?
Quartile	quartile of the dialogue the RE was uttered in
Dial_No	number of dialogues already completed +1
Mention_No	number of previous mentions of target +1

Table 3: The Alignment feature set.

Map Features	
Main_Map_type	most frequent type of LM on this map
Main_Map_other	other attribute if the most frequent type of LM
Mixedness	are other LM types present on this map?
Ink_Orderliness	shape of the ink blot(s) on the IF's map
Lmprop Features	
other_Att	type of the other attribute of the target
*[Att]*_Value	value for each *Att* of target
*[Att]*_Difference	was *Att* of target different between the two maps?
Missing	was target missing on one of the maps?
Inked_Out	was target inked-out on the IG's map?
Speaker Features	
Dyad_ID	ID of the participant-pair
Speaker_ID	ID of the person who uttered this RE
Speaker_Role	was the speaker the IG or the IF?

Table 4: The Ind feature set.

	Initial references	Subsequent references	Combined
training set	4,140	9,909	14,038
test set	1,796	4,296	6,103
total	5,936	14,205	20,141

Table 5: The sizes of training and test sets for the three data subsets.

of this feature.

3. **Incl_*Att*** records whether attribute *Att* has been included in the referring expression for all attributes that precede the current one in the chain. This feature captures an aspect of serial dependency that is not usually represented in traditional REG approaches: it allows the machine learner to pick up patterns of two attributes occurring together particularly often or rarely.

We built six chained models based on combinations of the chaining features, and one which included both all the chaining features and all the non-chaining features:

- **1**: uses only the DP_*Att* features.
- **2**: uses only the DP_RE features.
- **3**: uses only the Incl_*Att* features.
- **1+2**: uses the DP_*Att* and the DP_RE features. This model is most likely to emulate the behaviour of the IA or the Greedy Algorithm, if this behaviour is supported by the data.
- **2+3**: uses the DP_RE features and the Incl_*Att* features.
- **1+2+3**: uses all chaining features.
- **1+2+3+NC**: uses all chaining and all non-chaining features.

We used the C4.5 algorithm implemented in the Weka toolkit (Witten and Frank, 2005) to train the decision trees used in our models.

Results

For the experiments described here, we used a 70–30 split to divide the data into a training set and a test set. We performed separate tests for subsequent and initial references and for the complete data set. Table 5 shows the sizes of training and test sets for the three data subsets.

In addition to the main prediction class **content pattern**, the split was stratified for Speaker_ID and Quartile to ensure that training and test set contained the same proportion of descriptions from each speaker and each quartile of the dialogues.

	Initial references	Subsequent references	Combined
1	39.9%	43.7%	40.4%
2	42.0%	41.8%	38.5%
3	39.0%	41.4%	37.4%
1+2	42.3%	44.3%	41.5%
2+3	42.0%	41.8%	38.5%
1+2+3	42.9%	44.3%	41.4%
1+2+3+NC	72.5%	66.4%	68.6%
NC	72.3%	66.0%	68.2%

Table 6: Accuracy achieved by our models.

	Initial references	Subsequent references	Combined
1	39.9%	43.7%	40.4%
1-nc	39.9%	49.0%	46.0%

Table 7: Comparison of chained and unchained features representing the discriminatory power of the three attributes.

Table 6 shows the Accuracy achieved by our models in replicating the human-produced data. Accuracy records the proportion of content patterns that the models replicated perfectly. We tried all possibly orders in which the three attributes can be chained, but report only the result of the best performing order for each chained model.

The results show that the model based on non-chaining features only (NC) vastly outperforms all models using only chaining features on both initial and subsequent reference (χ^2(1+2,NC)=6646.5, df=1, p$\ll$.01). Adding the chaining features to the non-chaining features (1+2+3+NC) does not result in a significant improvement over the performance of the non-chained model (χ^2(NC,1+2+3+NC)=1.5, df=1, p$>$.2).

These results suggest that the serial dependency embodied in traditional REG algorithms is not a necessary feature of human production of referring expression generation; it appears that factors other than discriminatory power better explain the referring behaviour of human speakers in task-oriented dialogue.

Considering the similarity of the DP_*Att* features to the non-chaining Prop_Vis_Same_*Att* and Prop_Intervening_*Att* features, we also compared these two features directly by training a model that uses only Prop_Vis_Same_*Att* and Prop_Intervening_*Att* for each of the three attributes (1-NC). Table 7 shows the performance of model 1, based on the chaining feature DP_*Att*, and the model 1-NC based on the equivalent non-chaining features to those from model 1. This comparison shows that, for subsequent reference, using the non-chaining version of this feature outperforms the chaining version (χ^2=159.1, df=1, p$\ll$.01), while no difference is observed between the two feature sets for initial reference.

The non-chaining features Prop_Vis_Same_*Att* and Prop_Intervening_*Att* represent the discriminatory power of the individual properties of the target referent independently of the referring expression under construction. This is similar to the notion of visual salience of an attribute, which is usually taken to be determined by how different it is from those of the objects surrounding the intended referent. The results from Table 7 therefore indicate that visual salience might be of more importance in the choice of attributes for referring expressions than dynamically-computed discriminatory power.

Conclusions

In this paper we set out to explore whether the serially dependent manner in which most traditional REG algorithms choose attributes to include in referring expressions is evidenced in the referring behaviour of human dialogue participants. We used an approach based on machine-learned decision trees, one for each attribute of the target referent. We built two different types of models: chained models, in which some or all of the features available to each tree are dependent on the decisions that the previous trees in the chain have made; and non-chained models, where only features independent of the referring expression under construction were used.

Our results suggest that serial dependency does not play a significant role in human referring expression generation. The model based only on non-chaining features outperformed all models based only on chaining features, and combining chaining and non-chaining features in one model did not increase performance.

A direct comparison of features representing the discriminatory power of the individual attributes which was either pre-computed at the start (a simple non-chained model) or computed at the moment at which the attribute was under consideration and taking into account the already included attributes (a simple chained model) showed a slight advantage for the pre-computed version. This suggests that people do not compute the discriminatory power of an attribute when they decide whether to use it, but rather make a decision based on the attribute's visual salience within the visual context of the intended referent.

Overall, our results bring further support for our proposal in (Dale and Viethen, 2009) that the attributes in a referring expression might be chosen independently and in parallel, based on simple scene analysis rather than on more computationally expensive selection processes.

References

Brennan, Susan E. and Herbert H. Clark (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22:1482–1493.

Carroll, John M. (1980). Naming and describing in social communication. *Language and Speech* 23:309–322.

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

Dale, Robert (1989). Cooking up referring expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, 68–75. Vancouver BC, Canada.

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

Dale, Robert and Jette Viethen (2009). Referring expression generation through attribute-based heuristics. In *Proceedings of the 12th European Workshop on Natural Language Generation*, 58–65. Athens, Greece.

van Deemter, Kees (2006). Generating referring expressions that involve gradable properties. *Computational Linguistics* 32(2):195–222.

Gardent, Claire (2002). Generating minimal definite descriptions. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 96–103. Philadelphia PA, USA.

Gatt, Albert (2007). *Generating Coherent Reference to Multiple Entities*. Ph.D. thesis, University of Aberdeen, UK.

Goudbeek, Martijn and Emiel Krahmer (2010). Preferences versus adaptation during referring expression generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 55–59. Uppsala, Sweden.

Horacek, Helmut (2004). On referring to sets of objects naturally. In *Proceedings of the 3rd International Conference on Natural Language Generation*, 70–79. Brockenhurst, UK.

Jordan, Pamela W. and Marilyn Walker (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research* 24:157–194.

Krahmer, Emiel and Mariët Theune (2002). Efficient context-sensitive generation of referring expressions. In Kees van Deemter and Rodger Kibble (Eds.), *Information Sharing: Reference and Presupposition in Language Generation and Interpretation*, 223–264. CSLI Publications, Stanford CA, USA.

Krahmer, Emiel, Sebastiaan van Erk and André Verleg (2003). Graph-based generation of referring expressions. *Computational Lingustics* 29(1):53–72.

Louwerse, Max M., Nick Benesh, Mohammed E. Hoque, Patrick Jeuniaux, Gwyneth Lewis, Jie Wu and Megan Zirnstein (2007). Multimodal communication in face-to-face computer-mediated conversations. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 1235–1240.

Pickering, Martin J. and Simon Garrod (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences* 27(2):169–226.

Viethen, Jette, Robert Dale and Markus Guhe (2011). Generating subsequent reference in shared visual scenes: Computation vs. re-use. In *Proceeding the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, UK.

Witten, Ian H. and Eibe Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco CA, USA.